Reconstruction de profils protéiques pour la recherche de biomarqueurs

Pascal Szacherski

Numéro d'ordre: **4740**

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ BORDEAUX 1

École Doctorale **Sciences Physiques et de l'Ingénieur (ED209)**

préparée au laboratoire **Électronique et Systèmes pour la Santé (LE2S) du CEA Léti, Minatec Campus, DRT/DTBS/STD**

en collaboration avec le laboratoire d'**Intégration du Matériau aux Systèmes, Groupe Signal/Image**

par **Pascal SZACHERSKI**

pour obtenir le grade de

DOCTEUR

Spécialité : **Automatique, Productique, Signal et Image, Ingénierie Cognitique**

Reconstruction de profils protéiques pour la recherche de biomarqueurs

Directeur de thèse: **Jean-François GIOVANNELLI**
Co-directeur de thèse: **Pierre GRANGEAT**

soutenue le 21 décembre 2012

devant la commission d'examen formée de:

M. Daniel COMMENGES,	Professeur à l'Université Bordeaux 2,	Président
M. Jérôme IDIER,	Directeur de Recherche CNRS,	Rapporteur
M. Alfred HERO III.,	Professeur à l'Université du Michigan,	Rapporteur
M. Jean-Philippe CHARRIER,	Directeur de Recherche bioMérieux,	Examinateur
M. François CARON,	Chargé de Recherche INRIA Bordeaux,	Examinateur
M. Audrey GIREMUS,	Maître de Conférence, IMS Bordeaux,	Examinateur
M. Jean-François GIOVANNELLI,	Professeur à l'Université Bordeaux 1,	Directeur de thèse
M. Pierre GRANGEAT,	Directeur de Recherche CEA,	Co-directeur de thèse

Université Bordeaux 1
Les Sciences et les Technologies au service de l'Homme et de l'Environnement

Date de compilation du document : 21 janvier 2013.

Für meine Onkel Michael und Hans-Georg

Table des matières

Annexes

Remerciements

FELIX QUI POTUIT RERUM COGNOSCERE CAUSAS

(« Heureux celui qui a pu pénétrer le fond des choses. »)

Virgile, *Géorgiques*

L'aventure que je vais vous raconter et que nous allons re-vivre ensemble commença avec ce message, reçu un matin d'été 2009 au moment où je commençais une journée de stage de fin d'étude :

De :	Jean-François Giovannelli
À :	Pascal Szacherski
Date :	10 juillet 2009, 08h17
Sujet :	Sujet de thèse ...

```
Bonjour Pascal.

J'espère que vous vous portez bien et que votre stage répond à vos attentes.
En fait, je n'ai guère de doute qu'il se déroule bien et j'ai eu quelques
nouvelles par l'intermédiaire de Charles DOSSAL ces derniers temps.

Je reviens vers vous avec une proposition de thèse, même si je me souviens
que vous avez été réticent au printemps... Nous sommes à la recherche, avec
Pierre Grangeat (CEA-LETI, Grenoble), d'un doctorant pour travailler sur un
sujet en protéomique et signal. Ce sujet est disponible en suivant le lien
suivant : http://giovannelli.free.fr/Divers/TheseProteomique.pdf.

Il s'agit de travailler à des méthodes d'inversion fondées sur des
techniques de statistiques bayésiennes et d'échantillonnage stochastique (du
type de celles sur lesquelles vous avez travaillé en projet "Déconvolution
d'images").

Le travail se déroule essentiellement à Grenoble, avec un financement CEA
(type CIFRE). Il présente des aspects plus académiques, fondamentaux et des
aspects plus appliqués, industriels.

Au final, je pense que le travail présente un certain équilibre et une
certaine variété qui je l'espère retiendra votre attention.

En attendant votre réponse, que je continue d'espérer positive,
Bien cordialement,
JFG.
```

Depuis ce mail, ma vie a changé, et vous tenez une des preuves dans vos mains (à moins que vous l'affichiez sur l'écran de votre ordinateur, tablet et consortes).

Premièrement dans ma vie privée. Des aller-retours Grenoble ↔ Sarrebruck en train, le vendredi après-midi, le dimanche soir voire la nuit. Si les trajets ne me manquent pas, les escapades sarroises le font : discuter en se promenant le long de la Sarre, prendre une bière dans un parc, les samedi matins au marché local où la dame du boucher nous gâtait avec un bout de saucisse et le boucher avec un excellent pâté de foie (certains collègues se souviennent peut-être des sacs qui pendaient de ma fenêtre de bureau le lundi matin ... maintenant vous savez ce qu'ils contenaient), mais aussi l'annonce de la grossesse de

ma compagne, la joie, le bonheur, l'angoisse qu'elle ne puisse pas me rejoindre, puis le soulagement quand elle a eu sa mutation.

En mai 2011, à mi-parcours de thèse, notre fille est née. Quel moment inoubliable ! Difficile à décrire, tellement ça m'a pris … j'ai pleuré … je l'ai serrée dans mes bras … j'étais papa.

Un an est passé avec des progrès phénoménaux, l'installation en Chartreuse, le patrimoine, les travaux (avec l'aide de mes beaux-parents, merci !), les courgettes, les salades, le potager, les amis, la neige (liste non exhaustive dans le désordre) … et le vélo. (Parfois même pour prendre la navette aéroport à Voiron … certains s'en souviennent.)

Et puis, l'annonce un peu inattendue de notre deuxième grossesse… Le ventre s'arrondit, les premiers contacts, de plus en plus intenses, puis des coups, et bientôt la naissance !

Tout cela, je l'ai vécu avec ma compagne, Katia, à qui j'adresse mes plus tendres remerciements … pour son soutien, son amour, sa volonté de comprendre mon sujet de thèse malgré le fait d'être professeur d'allemand, les moments passés ensemble, les moments que nous passerons ensemble … **Merci** !

Dans cette début de carrière scientifique, j'ai rencontré beaucoup de monde, à commencer par mes directeurs de thèse *Jean-François* et *Pierre*. Je leur suis infiniment reconnaissant pour la patience profonde, les explications riches mais parfois incomprises, les discussions scientifiques, formelles et informelles à la machine à café de l'IMS où Jean-François a trop souvent payé pour moi, les visites et accueils enrichissants, les « trifouillages d'équations », les relectures innombrables, le soutien permanent, l'amitié, … bref, pour ces trois années dans la jungle protéo-bayésienne qui compte parmi les plus vicieuses peut-être ! **Merci !**

Merci aussi à tous ceux qui m'ont accompagné pendant cette aventure. Je n'avais pas l'intention de faire une liste, énumérer des noms, et surtout pas de les ordonner de quelle façon que ce soit. Finalement, j'ai dû le faire, et vous m'excuserez, j'espère, si l'ordre ne vous convient pas. Mes remerciements donc, dans le désordre structuré, aux personnes suivantes.

- *Rémi*, avec qui je suis toujours d'accord — au niveau bayésien au moins. Quel plaisir d'avoir discuté les fondements des aspects bayésiens (au moins ceux qui nous étaient accessibles …) autour d'une tasse de thé ou au tableau, souvent pointé d'une touche d'humour, mais aussi d'avoir fait quelques sorties de vélo avec toi. À deux, c'est quand même plus sympa ! Une dernière chose : je n'oublierai jamais 1982 … (par contre, Pirlo en 2006, j'ai déjà oublié !)
- *Abbas*, يا عبّاس شكرا جزيلا لالوقت معك في مكتبنا
  Le bureau des thésards, malgré ce qu'on peut croire, était très efficace, surtout en termes de Rrrrrrrrrrrrrrrronda. — Tu nous as fait découvrir un peu de ton pays, nous as fait rire … et peur (au moins deux fois).
- *Raphaëlle*, la maîtresse du casque EEG. Mais même si je ne t'ai pas convertie en bayésienne, tu maîtrises LaTeX maintenant ! Après la thèse comme pendant, Capt'n LaTeX est là pour toi ! (Et tu as vu ? J'ai mis les trémas ! À toi maintenant de mettre

les trémas sur le u de Tübingen. :-))
- *Andriy*, Андрий … je devrais t'appeler bouquetin car je n'ai jamais vu quelqu'un grimper les montagnes comme toi. Mais bon, un bouquetin sait pas non plus jouer au backgammon comme toi !
- *Laurent*, à travers nos discussions tu as partagé avec moi tes connaissances de la protéomique et ses applications, et parfois ton incompréhension par rapport à des sujets brisants. Ton implication dans le développement et dans le bon fonctionnement de cette thèse (et du projet BHI-PRO) est incontestable ! C'était très agréable d'être ton coéquipier !
- *Venceslass*, tu étais un peu mon parrain (peut-être ne le savais-tu pas ?) pendant les premiers mois : tu avais la réponse à toutes questions, tu avais toutes les clés (même pour ouvrir les portes donnant accès à mon compteur de gaz !). Après une dizaine de mois, j'ai dû quitter notre bureau commun, mais bon, tu restes attaché à la protéomique malgré toi !
- *Pascale*, tu as également partagé ton bureau avec moi pendant plus d'un an … et pourtant, je ne t'ai pas appris beaucoup d'allemand malgré ta volonté. Tu me pardonneras certainement … en échange contre une tablette de bon chocolat Ritter Sport que tu sauras ouvrir comme une grande ?
- Les stagiaires PROTIS *Riheb*, *Abrar* et *Abdoulaye* pour vos travaux qui m'ont inspiré et aidé.
- Le chef du laboratoire LE2S, *Régis*, pour sa confiance et tous les *membres du labo*, actuels ou anciens de ces trois dernières années, vous que j'ai croisés dans les couloirs ou au café, certains d'entre vous m'ont raconté les histoires d'antan ou d'aujourd'hui, ont discuté vélo, rando ou foot avec moi ou m'ont parlé en allemand (« parlé, lu, écrit »). — Même si j'ai vu mes collègues du *Groupe Signal Image de l'IMS* beaucoup moins souvent, je tiens à vous mentionner, notamment *Cornelia*, *Audrey* et *Yannick*.
- Que tout le monde le sache (si ce n'était pas déjà le cas) : nos *secrétaires* omniprésentes font tellement de choses pour nous mais sont **hélas** souvent négligées. Pensons à elles, et je le fais : votre présence à vous toutes, en particulier *Michelle* qui profite bien de sa retraite maintenant, m'a facilité indéniablement la vie professionnelle, et je vous en suis reconnaissant !
- *Dominique* et *Sébastien*, vous qui m'avez gentiment accueilli à maintes reprises quand j'ai fait des déplacements à Bordeaux et récemment à Paris. Même si c'était souvent pour peu de temps, j'ai toujours apprécié d'être avec vous.
- Le consortium de BHI-PRO, nous avons fait un beau bout de chemin ensemble, nous avons spécifié, convergé, discuté, rigolé à Lyon, Marcy, Dijon, Bordeaux, Saclay, Grenoble. C'était formidable de travailler avec vous, et je vous souhaite de bien terminer le projet, notre projet !
- Je m'en voudrais de passer sous silence les membres de mon jury de thèse, *Audrey Giremus*, *Jérôme Idier*, *Alfred Hero III.*, *Jean-Philippe Charrier*, *François Caron* et *Daniel Commenges* qui ont analysé dans les détails les fruits fraîchement cueillis de ces dernières années.

Pour finir, si vous avez encore un peu de souffle, laissez-moi exprimer encore deux « merci » inhabituels : Merci à tous les arbres de thés dont j'ai fait infusé les feuilles. Vous m'avez donné la force d'un chêne allemand dès le matin, parfois très tôt. Vous avez coloré l'eau dans laquelle je vous ai noyées, vous avez libéré la théine et le tanin dans la théière, et vous avez déployé toute votre saveur dans ma bouche.

Puis, merci aux auteurs de tous les livres qui m'ont accompagné pendant les nombreux voyages en train pour rejoindre compagne et patrie ainsi que pendant les trajets quotidiens en car pour voir mon bureau très souvent désordonné. Je ne pourrai pas les

citer tous, mais je propose aux curieux rats de bibliothèque une liste de recommandations de mes lectures ci-dessous.

Danksagung

Drei Jahre sind vergangen, in denen ich viel gelernt und erlebt habe. Viele lange Zug-Reisen nach Saarbrücken zu meiner Lebensgefährtin Katia, die nach etwa einem Jahr ein Ende genommen haben. Die Geburt unserer Tochter, und das unbeschreibliche, bewegende, stolze Gefühl, sie das erste Mal in meinen Armen willkommen zu heißen. Die korrekte Benutzung eines Akku-Schraubers für unser trautes Heim [↓1], der Anbau (und der Verzehr!!) unseres eigenen Gemüses, der Aufbau einer Schaukel, ... Und mit den gleichen Händen, mit denen ich all dies realisiert habe, mit denen ich diese Arbeit hier tippe, ja, mit diesen Händen spüre ich die Tritte (und den Schluckauf) unseres zweiten Kindes durch Katias Bauchdecke. Ich danke dir, Katia, zärtlich für dies, deine Unterstützung, dein Vertrauen, deine Zuversicht, und für noch so vieles mehr.

In diesem letzten Abschnitt des meistgelesenen Kapitels dieser Arbeit möchte ich mich zusätzlich bei denjenigen bedanken, die mich „auf deutscher Seite“ begleitet und unterstützt haben. Dieser Dank richtet sich insbesondere an meine Familie, die mir moralisch und auch finanziell beistanden, vom kleinen Ein-mal-Eins über das Studium im In- und Ausland bis zum heutigen Tag, den leider nicht mehr alle erleben können. Unter ihnen meine Onkel Michael und Hans-Georg, denen ich diese Arbeit widme.

Le Choix du Libraire

C. Beavan. *No Impact Man : Saving the planet one family at a time*. Piatkus, juillet 2011. ISBN 0749953209.

T. Ben Jelloun. *Cette aveuglante absence de lumière*. Points, janvier 2002. ISBN 2020530554.

A. Galland. *Les Mille et Une Nuits (Tomes 1, 2, 3)*. Flammarion, mai 2004. ISBN 2080712004, 2080712012, 2080712020.

E. Harding-Esch and P. Riley. *The Bilingual Family : A Handbook for Parents*. Cambridge University Press, 2ème édition, mars 2003. ISBN 0521004640.

P. Jaenada. *Le chameau sauvage*. J'ai lu, novembre 2005. ISBN 2290349534.

D. Keyes. *Les Mille et Une Vies de Billy Milligan*. Le Livre de Poche, janvier 2009. ISBN 2253125024.

D. Keyes. *Flowers for Algernon*. Harvest Books, mars 2010. ISBN 015603008X.

D. Pennac. *Au bonheur des ogres (et les autres romans de la septologie de Malossène)*. Folio, octobre 1997. ISBN 2070403696.

J. K. Rowling. *Harry Potter (Septologie)*. Bloomsbury, novembre 2010. ISBN 1408812525.

A. Smith and J. B. Mackinnon. *The 100-Mile Diet : A Year of Local Eating*. Vintage Canada, octobre 2007. ISBN 0679314830.

W. Szpilman. *Le Pianiste*. Pocket, janvier 2003. ISBN 2266117068.

A. Walker. *On Bicycles : 50 Ways the New Bike Culture Can Change Your Life*. New World Library, novembre 2011. ISBN 1608680223.

A.-X. Wurst. *Zur Sache, Chérie : Ein Franzose verzweifelt an den deutschen Frauen*. rororo, décembre 2010. ISBN 3499626144.

1. Ich hatte viel viel erfahrene Hilfe dabei ...

Conventions

Nous entreprendrons ensemble un voyage pendant lequel nous croiserons des distributions, des ensembles, des fonctions, des matrices, des vecteurs, des indices, peut-être quelques êtres morts, d'autres bien vivants. Avant le départ, je me permets de vous imposer « mes » notations que nous utiliserons tout au long de ce document. ↓2

Notations

$\mathbb{N}$	ensemble des naturels strictement positifs
$\mathbb{N}_0$	$\mathbb{N} \cup \{0\}$
$\mathbb{R}$	ensemble des réels
$\mathbb{R}_+$	ensemble des réels positifs
$\mathrm{card}(E)$	cardinalité de l'ensemble E
$\theta_{1:N}$	$[\theta_1, \ldots, \theta_N]$
x	scalaire
$\boldsymbol{x}$	vecteur
$\mathbf{A}$	matrice
$\mathbf{A}^{\mathrm{T}}$	transposée de la matrice $\mathbf{A}$
$\lvert x \rvert$	valeur absolue d'un scalaire
$\lVert \boldsymbol{x} \rVert_p$	norme p du vecteur $\boldsymbol{x}$
$\lvert \mathbf{A} \rvert$	déterminant d'une matrice carrée
$\mathrm{tr}(\mathbf{A})$	trace d'une matrice carrée
$\mathrm{diag}(\boldsymbol{x})$	matrice diagonale dont les éléments diagonaux sont les éléments de $\boldsymbol{x}$
$\Pr(X)$	loi de probabilité discrète de la variable aléatoire X
$\Pr(X = x)$	probabilité de l'événement $X = x$
$\mathrm{p}(x)$	densité de probabilité pour x
$\mathrm{p}(x \mid y)$	probabilité conditionnelle de x sachant y (*idem* pour le cas discret)
Ω_X	ensemble/intervalle/région de probabilité de la variable aléatoire X
$\mathrm{grad}(f(\boldsymbol{x}))$	gradient de la fonction f

Indices

$\mathsf{m} = 1, \ldots, \mathsf{M}$	classe
$n = 1, \ldots, N$	individu
$p = 1, \ldots, P$	protéine
$i = 1, \ldots, I$	peptide
$l = 1, \ldots, L$	fragment

Distributions

Loi normale La loi normale scalaire $\mathcal{N}(x; m, \gamma)$ pour le paramètre x est déterminée par sa moyenne m et sa précision γ (*i.e.* l'inverse de la variance), exprimée par

$$\mathcal{N}(x; m, \gamma) = (2\pi)^{-1/2} \gamma^{1/2} \exp\left(-\tfrac{1}{2}\gamma\,(x - m)^2\right).$$

2. Les notes en bas de pages sont indiquées dans le texte par un ↓ suivi d'un numéro afin de les distinguer d'éventuelles puissances en mathématiques.

La loi normale multivariée $\mathcal{N}(\boldsymbol{x};\boldsymbol{m},\boldsymbol{\Gamma})$ pour le vecteur de paramètres $\boldsymbol{x} \in \mathbb{R}^N$ est déterminée par son vecteur de moyenne $\boldsymbol{m} \in \mathbb{R}^N$ et sa matrice de précision $\boldsymbol{\Gamma} \in \mathbb{R}^{N\times N}$ (*i.e.* l'inverse de la matrice de covariance), exprimée par

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{m},\boldsymbol{\Gamma}) = (2\pi)^{-N/2}\,|\boldsymbol{\Gamma}|^{1/2}\exp\left(-\tfrac{1}{2}\,(\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}}\,\boldsymbol{\Gamma}\,(\boldsymbol{x}-\boldsymbol{m})\right).$$

Dans la suite, nous omettrons les termes « vecteur » et « matrice » dans l'utilisation de la loi normale quand ceci ne prête pas à confusion.

Loi uniforme La loi uniforme scalaire $\mathcal{U}(x;x^{\mathrm{m}},x^{\mathrm{M}})$ pour le paramètre x par rapport à l'intervalle $\left[x^{\mathrm{m}},x^{\mathrm{M}}\right]$ est donnée par

$$\mathcal{U}(x;x^{\mathrm{m}},x^{\mathrm{M}}) = \frac{1}{x^{\mathrm{M}}-x^{\mathrm{m}}}\mathbb{1}_{\left[x^{\mathrm{m}},x^{\mathrm{M}}\right]}.$$

La loi uniforme multivariée $\mathcal{U}(\boldsymbol{x};\boldsymbol{x}^{\mathrm{m}},\boldsymbol{x}^{\mathrm{M}})$ pour le paramètre $\boldsymbol{x} \in \mathbb{R}^N$ par rapport à l'intervalle multidimensionnel $\left[\boldsymbol{x}^{\mathrm{m}},\boldsymbol{x}^{\mathrm{M}}\right] = \left[x_1^{\mathrm{m}},x_1^{\mathrm{M}}\right]\times\cdots\times\left[x_N^{\mathrm{m}},x_N^{\mathrm{M}}\right]$ est donnée par

$$\mathcal{U}(\boldsymbol{x};\boldsymbol{x}^{\mathrm{m}},\boldsymbol{x}^{\mathrm{M}}) = \frac{1}{x_1^{\mathrm{M}}-x_1^{\mathrm{m}}}\cdot\cdots\cdot\frac{1}{x_N^{\mathrm{M}}-x_N^{\mathrm{m}}}\mathbb{1}_{\left[\boldsymbol{x}^{\mathrm{m}},\boldsymbol{x}^{\mathrm{M}}\right]}.$$

Loi gamma La loi gamma $\mathcal{G}(x;\alpha,\beta)$ pour paramètre x de forme α et d'échelle β est définie par l'équation

$$\mathcal{G}(x;\alpha,\beta) = \frac{x^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}\exp(-\,x/\beta\,),$$

$\Gamma(\cdot)$ désignant la fonction gamma.

Loi normale-wishartienne La distribution normale-wishartienne notée $\mathcal{NW}(\boldsymbol{m},\boldsymbol{\Gamma};\boldsymbol{\mu},\boldsymbol{\Lambda},\eta,\nu)$ pour le couple de paramètres $(\boldsymbol{m},\boldsymbol{\Gamma}) \in \mathbb{R}^N\times\mathbb{R}^{N\times N}$, paramétrée par la moyenne $\boldsymbol{\mu}$, la matrice d'échelle $\boldsymbol{\Lambda}$, le nombre d'échantillons η et de ν degré de liberté, s'écrit

$$\mathcal{NW}(\boldsymbol{m},\boldsymbol{\Gamma};\boldsymbol{\mu},\boldsymbol{\Lambda},\eta,\nu) = \frac{|\boldsymbol{\Lambda}|^{-\nu/2}}{2^{\nu N/2}(2\pi)^N\Gamma_N(\nu/2)}\,|\boldsymbol{\Gamma}|^{(\nu-N)/2}\cdot\exp\left(-\tfrac{1}{2}\left[\mathrm{tr}\left(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\right)+\eta\,(\boldsymbol{\mu}-\boldsymbol{m})^{\mathrm{T}}\,\boldsymbol{\Gamma}\,(\boldsymbol{\mu}-\boldsymbol{m})\right]\right)$$

où N est la dimension, $\Gamma_N(\cdot)$ est la fonction Gamma N-dimensionnelle définie par l'expression $\Gamma_N(x) = \pi^{N(N-1)/4}\prod_{p=1}^{N}\Gamma\left(x+(1-p)/2\right)$, et $\mathrm{tr}(\cdot)$ l'opérateur trace d'une matrice carrée.

... et une dernière chose ...

Ce document a été généré avec LaTeX et un certain nombre de paquets très utiles ; l'ensemble des graphes a été créé avec matlab2tikz de Nico Schlömer. La compilation du document a été effectuée avec la distribution MiKTeX 2.9 (Windows) et a été testée avec la distribution TeXLive 2012 (Linux). J'ai testé l'affichage correct du document avec les logiciels suivants : SumatraPDF (Windows, version 1.5+), Adobe Acrobat Reader (Windows, versions 9+), FoxitReader (Windows, versions 4+), PDF-XChangeViewer (Windows, version 2.5), MuPDF (Windows et Linux, version 1.1), Evince (Linux, version 3.4.0), Zathura (Linux, version 0.0.8.5), apvlv (version 0.1.1), Okular (Linux, version 0.14.3).

1 Introduction

Dans le monde de la recherche protéomique, les équipes font des efforts considérables pour développer des chaînes d'analyse basées sur la spectrométrie de masse pour la quantification de protéines et pour la découverte de biomarqueurs, *i.e.* des protéines permettant de différencier des états cellulaires. Les échantillons biologiques utilisés pour ce faire sont des mélanges complexes, et les variabilités induites à la fois par cette complexité et par les instruments de mesures sont importantes.

À travers cette thèse et en synergie avec le projet ANR BHI-PRO, nous proposons d'étudier des méthodes statistiques — issues du Traitement du Signal et de l'Information — qui sont adaptées afin de maîtriser les variabilités technologiques et biologiques et de rendre les traitements numériques plus robustes. Nous les inscrivons dans la démarche « problème inverse ». Il est d'abord nécessaire de décrire « mathématiquement » la chaîne d'analyse à la fois pour la « quantification » et pour la « classification » (deux termes que l'on définira dans ce document) ; l'inversion du modèle nous permettra de calculer des estimations de la quantité ou de la classe selon le problème posé.

Les ouvrages suivants constituent nos références principales dans ce manuscrit, et nous y référons à plusieurs reprises :

1. *Reconstruction de profils moléculaires: modélisation et inversion d'une chaîne de mesure protéomique*, [1] : la thèse de Grégory Strubel constitue la base du sujet traité ici. Elle concerne la modélisation mathématique du couplage « Spectrométrie de masse/ Chromatographie liquide » ainsi que la quantification de marqueurs initialement connus et prédéfinis dans un cadre bayésien.

2. *The Bayesian Choice*, [2] : ouvrage d'enseignement approfondi des statistiques bayésiennes, écrit par Christian Robert, un des « bayésiens » les plus connus. Il donne une vue détaillée sur le traitement de l'information bayésien et sur ses aspects numériques.

3. *Bayesian Data Analysis*, [3] : ouvrage complémentaire au précédent, avec un accès plus applicatif. La plupart des exemples sont issus des sciences politico-statistiques, les auteurs – Andrew Gelman et collaborateurs – faisant leurs recherches dans ce domaine.

Nous réunissons dans ce document, avec l'expérience des trois années de préparation de thèse, plusieurs disciplines dont les principales sont la *protéomique* et le *traitement du signal bayésien dans le cadre des problèmes inverses*. Le présent chapitre introduit le premier de ces deux aspects sans pouvoir être exhaustif ou remplacer un cours de protéomique. Nous décrivons également l'activité de l'équipe d'accueil PROTIS au sein du laboratoire LE2S et l'objectif du projet ANR BHI-PRO.

1.1 Protéomique

À travers les pages suivantes, nous montrons ce qu'est la protéomique, où elle est utilisée et comment. La protéomique est l'étude du protéome qui est l'ensemble des protéines d'un organisme [100]. Avant de définir ces derniers termes, notons que la protéo-

mique – contrairement à la génomique qui est également une étude moléculaire, mais beaucoup plus ancienne et avancée – est encore dans son enfance.

Parmi les activités de la protéomique, mettons en relief les suivantes.

Identification L'identification des protéines répond à la question : quelles protéines sont présentes dans un échantillon biologique donné ?

Quantification Le but de la protéomique quantitative est de déterminer les concentrations des protéines dans un échantillon biologique.

Analyse différentielle Liée aux aspects précédents est l'analyse différentielle des protéines répondant à la question suivante : quelles sont les protéines qui ont une expression différentielle quand on compare deux classes d'échantillons (sain/malade, traitement médicamenteux réussi/non réussi) ? Cette différentiation peut se faire au niveau de la présence ou absence d'un groupe de protéines, ou bien au niveau de la quantité différente selon la classe considérée.

Interaction Les protéines peuvent interagir les unes avec les autres. Dans certains cas, ce n'est même pas la protéine que l'on mesure mais son interaction avec une autre protéine ou un groupe de protéines. On parle dans ce cas de mesure indirecte de la protéine permettant d'en déduire ses propriétés.

Les protéines cibles caractéristiques du phénomène que l'on étudie – que ce soit pour l'identification, la quantification, l'analyse différentielle et l'interaction entre protéines – sont appelées *biomarqueurs*.

Ces quatre applications ont un facteur commun : les mesures à partir desquelles on tire des conclusions sont soumises à de fortes incertitudes. Le but de cette thèse est d'arriver à les maîtriser le plus possible, ce qui sera démontré dans la suite du document.

Après cette brève introduction sur la protéomique, et ayant parlé de *protéines* et de *protéome*, prenons le temps de les aborder en peu plus en détail.

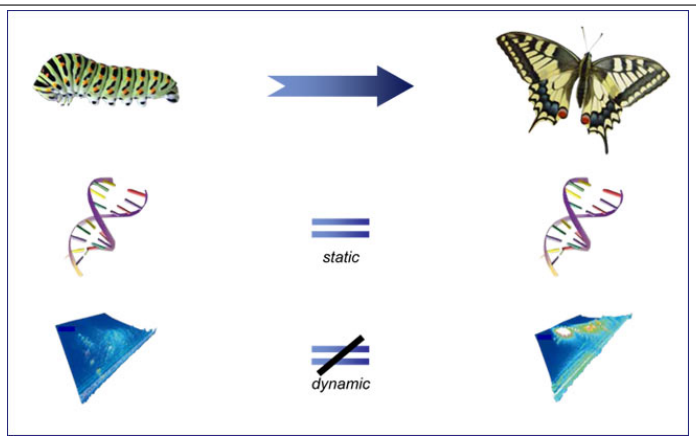
1.1.1 Protéines

Les protéines sont synthétisées par les cellules des organismes vivants par un mécanisme de transcription à partir des gènes [4]. Elles constituent donc l'expression cellulaire des gènes et sont les molécules de la vie. Certaines sont incluses dans des muscles ou contractent ces derniers, d'autres – les hormones – sont transmetteurs de signaux inter- et intracellulaires. Elles définissent également l'allure et la forme de cellules. De plus, les protéines contrôlent le système immunitaire par l'identification de germes et d'autres substances étrangères. Des protéines appelées enzymes contrôlent les réactions chimiques à l'intérieur des cellules, mais elles peuvent aussi être utilisées pour modifier d'autres protéines. Pour clore cette liste non exhaustive des fonctions, les protéines transportent de l'oxygène, du sucre, des substances nutritives, des déchets, et aussi d'autres protéines dans les cellules.

Tout aussi variée que leurs fonctions est la structure d'une protéine. Elle est constituée d'une longue chaîne composée d'acides aminés (AA). Il en existe vingt, et ils sont notés par leurs noms savants, des abréviations, ou simplement par des lettres de l'alphabet latin. La longueur de la chaîne d'AA d'une protéine est variable d'une protéine à l'autre. Ainsi, la protéine de l'insuline humaine a 110 AA et est relativement courte, la protéine *Neuron Specific Enolase* (NSE) compte 433 AA, et la protéine la plus longue est la titine avec 34 350 AA.

Notons enfin que les protéines sont créées par les gènes, mais la concentration des protéines, ce qui construit le *profil protéique*, est influencée par la cellule et son environnement.

Fig. 1.1 Le génome de la chenille et du papillon est le même, leur protéome ne l'est pas. Illustration issue de [101].



1.1.2 Protéome

Le protéome est l'ensemble de toutes les protéines d'un organisme, d'une cellule ou d'un échantillon biologique. Par exemple, le protéome humain est l'ensemble des protéines humaines.

Tandis que le génome est déterminé uniquement par héritage, le protéome, lui, est aussi déterminé par l'environnement. Autrement dit, le génome d'un organisme est statique, il ne change pas au cours du temps. Le protéome par contre est dynamique : au fil des années, le profil protéique change ce qui est dû à un certain nombre de facteurs, par exemple une maladie.

En effet, il y a des maladies qui sont purement héréditaires, donc génétiques comme par exemple le syndrome de Down, d'autres sont purement environnementales comme le scorbut. Beaucoup de maladies sont cependant entre les deux. Tandis que la génomique contribue grandement au traitement des maladies génétiques, la protéomique peut contribuer sur toute la gamme de ces maladies puisque les protéines sont créées à partir des gènes et les concentrations déterminées par l'environnement.

Pourquoi utilise-t-on le protéome plutôt qu'une protéine ? Classiquement, les études cliniques ont considéré une seule protéine. Cependant, le diagnostic d'une maladie dans ce mode de fonctionnement peut être délicat. Si une seule protéine n'est pas indicateur spécifique d'une maladie, on peut espérer qu'un ensemble de protéines l'est. La recherche à l'intérieur du protéome, considérant un profil protéique, permet d'étudier plusieurs protéines à la fois, mais aussi leurs dépendances, leurs corrélations. Un des résultats du projet européen LOCCANDIA [5] (Lab-On-Chip based protein profiling for CANcer DIAgnosis) souligne cette tendance : alors que les groupes « sain » et « cancer du pancréas ou pancréatite » n'ont pu être séparés à l'aide d'une protéine (MRP8/14 **ou** PAP1), il a suffi d'utiliser les deux protéines (MRP8/14 **et** PAP1) conjointement pour pouvoir séparer les **deux** groupes « sain » et « cancer du pancréas » ou « pancréatite » [6].

1.2 Domaine d'application de la protéomique

Les applications sont fortement liées aux activités de la protéomique (identification, quantification, interaction).

La protéomique a bien sûr un intérêt **scientifique**. Elle a été utilisée pour étudier des questions basiques de la biologie pour découvrir les structures des cellules, mais aussi

pour cartographier les protéines présentes dans un organisme.

Le protéome est dynamique, notamment par rapport à des maladies. Le **diagnostic** d'une maladie est une application très intéressante et très discutée dans la communauté. En effet, beaucoup de travaux traitent la question de la recherche de biomarqueurs pour des maladies comme le cancer, l'infarctus, et autres, dans le but de donner un diagnostic, et ce le plus tôt possible.

L'intérêt de la protéomique porte également sur le **suivi thérapeutique**, établissant des connaissances sur les distributions des biomarqueurs au fil des années et des étapes d'une maladie, ou sur la réaction du patient par rapport aux traitements.

Finalement, le **développement de médicaments** est également dans le champ applicatif de la protéomique. Comme la plupart des médicaments visent des protéines ou sont eux-mêmes des protéines thérapeutiques, il est naturel d'utiliser la protéomique pour identifier des médicaments candidats pour une maladie donnée.

1.3 Reconstruction de profils moléculaires : l'équipe « PROTIS »

Ce travail est réalisé sur le programme « laboratoire sur puce » en synergie avec le développement de composants spécifiques étudiés dans d'autres services du Laboratoire d'Électronique et de Technologie de l'Information (Léti). Ces activités relèvent du domaine des *Micro-* ou *Nano-Electro-Mechanical Systems* (MEMS/NEMS) appliqués à la biologie et à la santé. Un exemple est le projet européen LOCCANDIA sur lequel des micro-colonnes de digestion et de chromatographie intégrées sur silicium ont été étudiées [6]. Ces activités sont aussi réalisées en collaboration avec le Laboratoire d'Étude de la Dynamique du Protéome (EDyP) de la Direction des Sciences du Vivant (DSV) du CEA Grenoble.

L'équipe PROTIS (PROtéomique et Traitement de l'Information pour la Santé) à l'intérieur du laboratoire LE2S (Laboratoire Électronique et Systèmes pour la Santé) du CEA Leti inscrit ses travaux dans le traitement de l'information moléculaire. Son but est la reconstruction de profils moléculaires. Les recherches portent sur trois thématiques principales : l'exploitation des informations apportées par les données, l'apport des connaissances *a priori* par la proposition d'un modèle mathématique des phénomènes physiques en jeu, et finalement le développement d'algorithmes d'inversion pouvant intégrer les variabilités biologique et technologique.

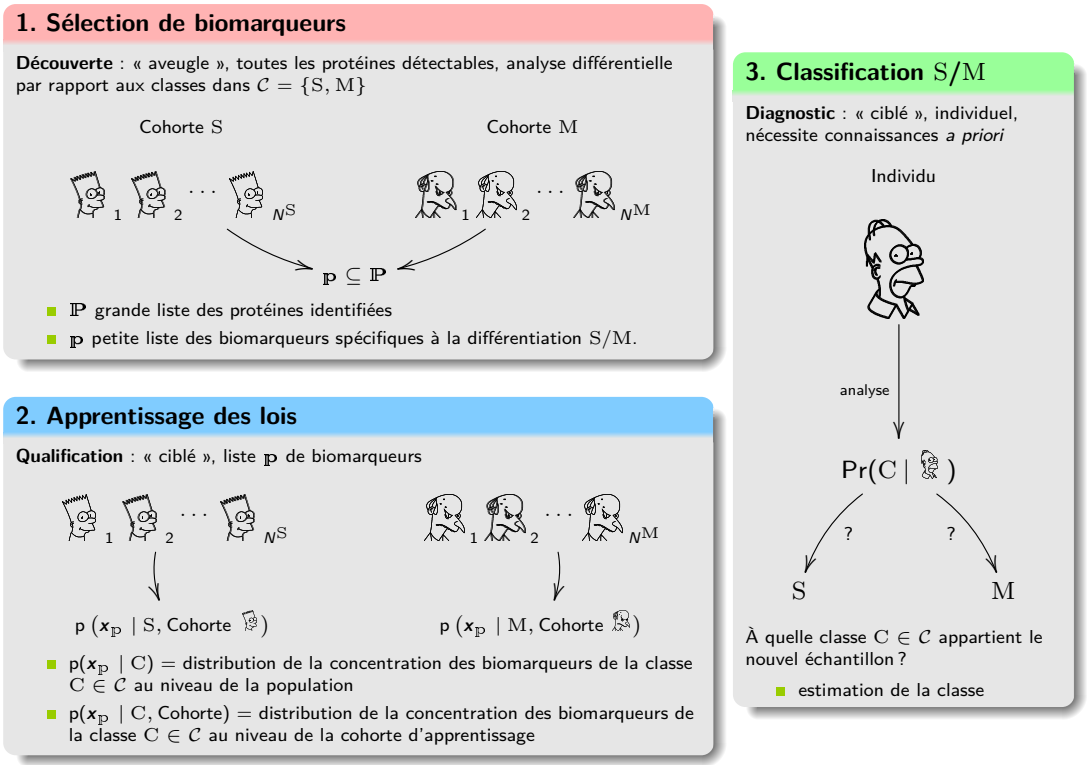
Parmi les composants utilisés, on trouve des capteurs nano-métriques NEMS qui sont sensibles à l'ajout d'une masse moléculaire. Ces composants permettent d'envisager le développement de nouveaux spectromètres de masses reposant sur la détection et la quantification de la masse sur des molécules individuelles, alors que les dispositifs actuels reposent sur une mesure collective. Le traitement et l'analyse des données acquises par un tel système est le sujet de thèse d'un membre de l'équipe, Rémi Pérenon [7, 8].

Dans cette thèse, nous nous intéressons plus particulièrement au couplage de la chromatographie liquide et de la spectrométrie de masse. À partir des données acquises, les questions que nous avons été amenés à étudier se résument aux suivantes : quantification de protéines comme base des développements, puis sélection de biomarqueurs, apprentissage des paramètres des classes et classification comme aide au diagnostic, cf. Fig. 1.2.

1.3.1 Sélection de biomarqueurs

Considérons un exemple concret. Nous supposons avoir à notre disposition une cohorte à deux classes : une classe d'individus atteints d'une maladie M, l'autre classe d'in-

Fig. 1.2 Les trois étapes majeures dans la reconstruction de profils moléculaires pour la découverte de biomarqueurs et la classification.



dividus « sains »[1] S. Nous effectuons une analyse différentielle pour savoir quel est le sous-protéome qui permet de différentier les deux classes. Autrement dit, quelle est la petite liste $\mathbb{p}$ de protéines biomarqueurs à l'intérieur de la grande liste $\mathbb{P}$ de toutes les protéines dont l'expression est différentielle ? Vu en termes d'apprentissage statistique, quelles sont les variables, inscrites dans $\mathbb{p} \subseteq \mathbb{P}$, qui permettent la distinction des classes ? Pour cette étape, on utilise généralement des méthodes non ciblées dû au grand nombre de protéines à observer qui permettent certes l'identification mais pas la quantification d'une protéine. C'est la raison pour laquelle nous avons besoin, après l'identification, d'une étape de connaissance des caractéristiques ou des paramètres des classes, comme par exemple la distribution de la concentration des biomarqueurs.

1.3.2 Apprentissage des paramètres des classes

Avec la connaissance acquise sur les biomarqueurs qui permettent de différentier deux classes, nous abordons ensuite le problème de l'apprentissage des paramètres des classes dans une approche ciblée. Notre équipe voit ce problème sous l'angle de la connaissance de la distribution de la concentration des biomarqueurs de chaque classe, $\mathrm{p}(\boldsymbol{x}_{\mathbb{p}} \mid \mathrm{B})$. Pour cela, nous avons à nouveau à notre disposition une cohorte avec les mêmes classes. (Il peut s'agir de la même cohorte ou d'une nouvelle, tant qu'il s'agit de la même maladie.) Ensuite, à l'aide de méthodes d'apprentissage statistique, nous déterminons les distributions dans chacune des classes [113, 114]. Cet apprentissage est fait dans l'objectif

1. *Stricto sensu*, il n'y a pas de classe d'individus sains, mais plutôt une classe d'individus non atteints de la maladie en question, ne montrant pas les symptômes spécifiques. Nous utilisons cependant cet abus de langage dans tout le document par soucis de simplicité.

de donner un diagnostic « sain – malade » lors de l'étude d'un nouvel échantillon. Ceci est le sujet de l'étape de classification suivante.

1.3.3 Aide au diagnostic par classification

Connaissant les biomarqueurs et les paramètres des classes, nous avons tous les prérequis pour résoudre un problème de classification. Les protéines à tester étant connues et bien identifiées, nous nous restreignons à celles-ci, d'où l'utilisation d'approches ciblées. En effet, à la réception d'un nouvel échantillon biologique à état biologique inconnu, nous pouvons désormais classer cet échantillon dans la classe d'individus malades M ou dans la classe d'individus sains S, donnant ainsi une information sur le diagnostic ou une aide au diagnostic [115, 116].

1.4 Inversion Hiérarchique Bayésienne : le projet « BHI-PRO »

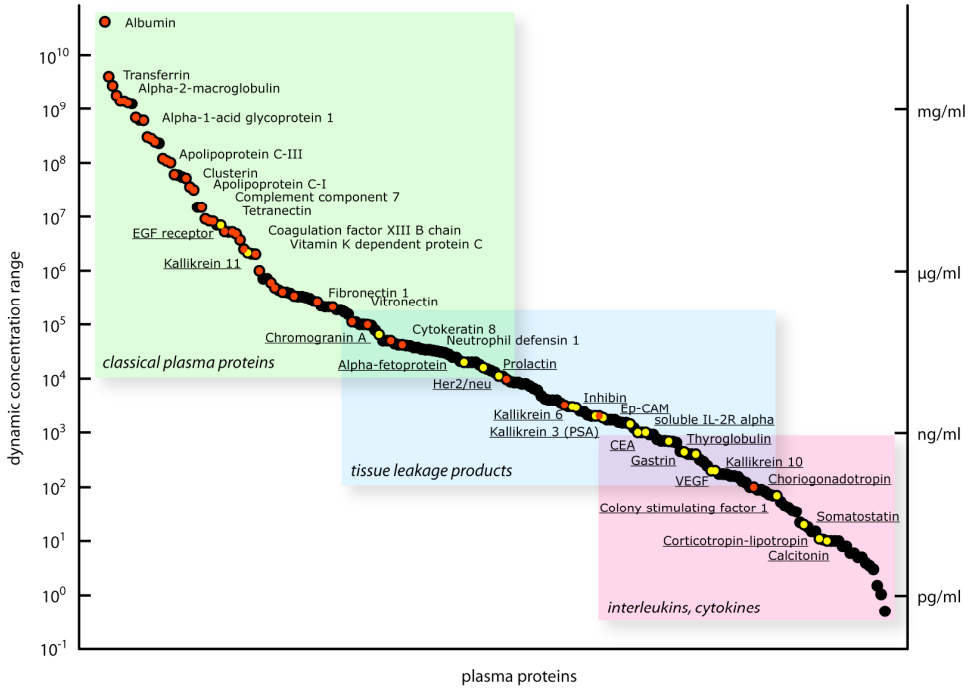
Cette thèse a été partiellement financée par le projet ANR BHI-PRO (contrat ANR-2010-BLAN-0313). Comme le démontre le résumé du projet ci-dessous, la thèse est en parfaite synergie avec les objectifs du consortium.

> *« Des efforts de recherche importants sont consacrés au niveau mondial pour développer des chaînes d'analyse reposant sur la spectrométrie de masse pour la découverte, la validation et la quantification de biomarqueurs protéiques dans des matrices complexes comme l'urine ou le sang. Cependant, maîtriser la variabilité technologique sur ces chaînes d'analyse est un point critique. Ceci nécessite de développer des techniques de traitement de l'information adaptées pour prendre en compte la complexité du mélange analysé, pour améliorer la fiabilité des mesures et pour faciliter l'usage de ces technologies.*
>
> *Une chaîne d'analyse protéomique est un enchaînement de traitements moléculaires qui peuvent être décrits par une structure de graphe, chaque nœud représentant un niveau d'analyse dans la chaîne. Chaque branche correspond à une décomposition moléculaire définissant un modèle de mélange hiérarchique. Dans ce projet BHI-PRO, nous proposons d'introduire des modèles hiérarchiques dédiés pour décrire les chaînes d'analyse MALDI et SRM/MRM[3]. Les nouveaux algorithmes d'inversion hiérarchique bayésiens reposeront sur deux innovations : l'association protéomique – problèmes inverses d'une part et problèmes inverses – échantillonnage stochastique d'autre part. La stratégie proposée repose sur des approches statistiques bayésiennes et des algorithmes d'échantillonnage stochastique.*
>
> *D'un point de vue biostatistique, la possibilité de tester plusieurs biomarqueurs simultanément fait partie des avantages de la protéomique. Cependant, quand le nombre de variables augmente, la probabilité de trouver des résultats par chance devient statistiquement significative. Nous proposons d'évaluer la puissance statistique des tests de discrimination dans le contexte bayésien étudié.*
>
> *Les principaux livrables seront deux logiciels d'inversion hiérarchique bayésien dédiés respectivement aux acquisitions MALDI et MRM, et un rapport de recommandations biostatistiques. » [9]*

1.5 Problématique de la thèse

À travers ses mécanismes internes, une cellule émet des protéines synthétisées à partir des gènes. L'étude de ces protéines permet de comprendre et de caractériser son fonc-

Fig. 1.3 Diagramme montrant le rapport dynamique des protéines du plasma humain. Les points jaunes représentent une protéine biomarqueur, majoritairement présentes dans la gamme inférieure. Figure issue de [10].



tionnement et son état. D'un point de vue biologique, le problème s'inscrit dans l'étude de marqueurs moléculaires, plus précisément de marqueurs protéiques. Pour l'utilisateur (chercheur en laboratoire protéomique, médecin, ...), il s'agit de disposer de méthodes de traitement numérique pour sélectionner, identifier, détecter et quantifier les marqueurs, puis éventuellement de classifier ou diagnostiquer à partir de ces derniers. Ceci peut concerner un marqueur unique, un ensemble de marqueurs ou même un réseau structuré de marqueurs. Cependant, les marqueurs sont souvent présents dans une très faible quantité comme montre la Fig. 1.3.

Un progrès majeur consisterait à réussir à quantifier ces marqueurs, et ce dans un milieu très complexe avec une grande dynamique et de fortes perturbations. Ces perturbations peuvent provenir des erreurs de mesures, d'incertitudes sur les paramètres, de l'imperfection des appareils sur le plan technique, ou bien de la variabilité biologique du taux de protéines d'un individu à l'autre.

Ceci justifie le recours à des modèles et des méthodes d'analyse relevant des approches statistiques. L'interprétation « propre » des mesures nécessite la modélisation de la chaîne d'acquisition ainsi que des techniques d'identification et d'étalonnage des paramètres. Une modélisation hiérarchique combine les variabilités technique et biologique à travers des paramètres en leur assignant un niveau spécifique. Le défi est ensuite d'estimer conjointement les paramètres, de prendre en compte les étalons et de réduire les variabilités.

Cependant, nous n'avons pas d'accès direct aux concentrations protéiques, nous devons passer par une version transformée, amplifiée, distordue du paramètre d'intérêt qui devient un paramètre caché. Donc, nous ne mesurons qu'*indirectement* ce que nous avons l'intention de mesurer, la mesure indirecte intégrant les variabilités diverses.

Exemple ⟨1.1⟩ (Mesures indirectes) *Donnons deux exemples de mesures indirectes qui sortent du cadre de la protéomique. Le but de la super-résolution est l'obtention d'une image haute résolution par une interpolation « intelligente » d'une image de basse qualité de référence à partir d'une série d'images, également de basse qualité. On part du principe que l'image à super-résoudre est une version échantillonnée, floutée et décimée de l'image de haute-qualité d'intérêt [11].*

En tomographie, on acquiert une série de mesures déportées à l'extérieur de l'objet d'intérêt. Dans le cas d'un scanner circulaire (type Tomographie par Émission de Positons, ou Tomodensitométrie), ces acquisitions sont des projections, réalisées sous des angles différentes ; on observe la transformée rayon X *(équivalente à la* transformée de Radon *en 2D) de l'objet [12, Ch. 12], [13].*

L'estimation peut être comprise comme la résolution d'un problème inverse. La réalité a fourni, en passant par des instruments de mesure, une observation. À partir de l'observation, comment retrouver la réalité ?

L'approche bayésienne propose un cadre formel adapté à l'estimation dans des cas complexes. Elle permet notamment de prendre en compte entre autres les modèles instruments, les informations *a priori* disponibles sur les paramètres et des modèles de bruits. Sa popularité dans la protéomique est prouvée par de nombreux ouvrages et articles comme par exemple la sélection [14–18].

Dans cette thèse, nous nous consacrerons au couplage des méthodes problèmes inverses et traitement statistique bayésien pour inverser la chaîne analytique décrite par la question de la **quantification** et de la **classification**, à partir de marqueurs très faiblement abondants. Le travail repose partiellement sur la thèse [1] où l'axe de recherche était centré sur l'estimation conjointe des paramètres instruments et des concentrations de protéines cibles, appliquée aux données obtenues par le couplage de chromatographie liquide et spectrométrie de masse. Le nouveau travail franchit une étape importante à la fois en termes de retombées pratiques et en termes de technologie de traitement : il s'agit d'étendre l'approche développée précédemment aux problématiques où l'on doit comparer des familles d'échantillons et attribuer un nouvel échantillon à une de ces familles. Ladite comparaison permettra également d'identifier et quantifier les protéines qui s'expriment de manière différentielles entre les familles étudiées.

Pour ce faire, cette thèse proposera une modélisation « directe » du problème en décrivant la structure hiérarchique de la chaîne d'analyse. En plus de l'étage « observations », cette hiérarchie comporte notamment des étages « technologique », « bio-technique » et « biologique » pour la quantification ; nous proposons ensuite de l'étendre en introduisant un étage « classe/paramètres de classe » pour la classification. Ceci nous demande de passer d'une méthode d'estimation de paramètres continus (quantité de protéines) à une méthode d'estimation de paramètres discrets (classe d'appartenance). En effet, comme nous allons le voir, cette extension s'intègre de manière immédiate dans les développements grâce à la souplesse de ce type de modélisation. L'inversion par des méthodes statistiques bayésiennes bénéficiera grandement de cette modélisation hiérarchique.

1.6 Structure du document

Ce manuscrit de thèse est structurée comme suit.

Ch. 2 Raisonnement bayésien. La thèse s'inscrit dans le cadre des méthodes statistiques bayésiennes pour le traitement du signal et la résolution des problèmes inverses. Nous proposons dans ce chapitre quelques notions de bases pour la compréhension et l'appréciation des développements qui suivront.

Ch. 3 **Modélisation physique et probabiliste de la chaîne d'analyse.** Nous modélisons la chaîne d'analyse de la préparation jusqu'à l'obtention du signal. Pour cela, nous proposons des modèles directs physiques et probabilistes pour deux modes d'acquisition de données, exposés dans ce chapitre.

Ch. 4 **Inversion-Quantification.** Nous nous consacrons à la quantification de protéines, en mettant le focus sur le mode d'acquisition *Selected Reaction Monitoring*. Pour ce faire, nous développons l'inversion permettant d'estimer la concentration protéique et décrivons sa mise en œuvre numérique. Les performances de l'Inversion-Quantification sont évaluées à partir de données simulées, réelles synthétiques et réelles cliniques.

Ch. 5 **Inversion-Classification.** Parmi les points forts de cette thèse compte indéniablement les travaux de l'Inversion-Classification. Nous l'exposons dans le Ch. 5, justifions l'intérêt du couplage de l'inversion et de la classification, détaillons les développements théoriques et numériques. Nous terminerons le chapitre sur des résultats sur des données simulées et sur des données cliniques.

Ch. 6 **Conclusions et perspectives.** Le corps de cette thèse se termine par la conclusion des travaux effectués. Nous donnons les perspectives pour la poursuite des développements de la reconstruction de profils moléculaires.

Nous veillons à la lisibilité et la légèreté du corps du texte. Certains détails qui ne sont pas d'intérêt immédiat y sont omis et présentés en annexe. Nous proposons ainsi au lecteur les suppléments suivants.

Ann. A **Calculs détaillés.** Le lecteur intéressé par les calculs détaillés et les preuves trouve son compte dans ce chapitre annexe.

Ann. B **Calculs de lois.** Le calcul des lois *a posteriori* (conditionnelles) est souvent assez immédiat ; cependant, les écrire en détail peut être relativement long. Afin d'éviter une surcharge gênante dans le corps du texte, nous les avons rapportés dans l'annexe.

Ann. C **Choix de modèle.** Les méthodes de choix de modèle ont été évoquées lors des entretiens de travail à Bordeaux. Nous proposons de revoir les notes dans ce chapitre, même si elles n'ont pas été utilisées dans le développement de l'Inversion-Classification.

Ann. D **Résultats de quantification LC-Full-MS.** Ayant mis le focus dans le chapitre Inversion-Quantification sur le mode d'acquisition SRM, les résultats obtenus pour la quantification avec l'instrument considéré en début de thèse sont présentés et commentés dans ce chapitre.

Ann. E **Notes biographiques.** Le dernier chapitre d'annexe résume quelques notes biographiques des personnages dont on ne connaît parfois que le nom. Le renvoi est indiqué par un « [i] » dans le corps du texte.

Finalement, nous terminons le document avec les références, une bibliographie personnelle et l'index des mots clés.

2 Raisonnement bayésien

Dans ce chapitre, nous introduisons le raisonnement bayésien et le calcul bayésien en nous concentrant sur les bases nécessaires pour comprendre la suite, les développements théoriques et algorithmiques du présent document. Le lecteur intéressé trouvera dans les ouvrages [19] [1], [2] et [3] des éléments théoriques, mais aussi pratiques (complété entre autre par les ouvrages [12, 20, 21]) qui vont plus loin dans les détails, souvent bien plus loin que ce dont nous avons besoin.

Si souvent un livre sur les statistiques bayésiennes est dédié aux approches bayésiennes et un livre sur les statistiques fréquentistes dédié aux approches fréquentistes, l'ouvrage [22] est un cours complet introduisant les statistiques du point de vue des deux approches, montrant qu'il y a des similitudes sous certaines conditions. Le cours [23] donne également une bonne introduction quant à ce point. Nous ne discuterons pas ici les convergences et divergences entre « fréquentistes » et « bayésiens ». Cependant, Tab. 2.1 résume rapidement les différences principales.

2.1 L'inversion dans un cadre bayésien

Quand une quantité physique n'est pas directement mesurable, on se sert d'autres quantités qui sont liées à la première et qui sont observables [12, 25]. L'estimation de la quantité d'intérêt doit passer par l'exploitation efficace des observations. La notion de *problème inverse* correspond à l'idée d'inversion de la chaîne d'acquisition pour accéder indirectement à la quantité d'intérêt.

Dans la littérature, beaucoup de méthodes d'inversion ont été décrites (*cf.* par exemple les deux ouvrages cités ci-dessus). Le choix d'une méthode dépend naturellement aussi de la physique du problème. Dans notre cas, nous disposons de mesures indirectes des concentrations protéiques, la chaîne analytique est perturbée par des processus aléatoires et de fortes variabilités, et les pics caractéristiques sont noyés dans du bruit de mesure. L'exploitation efficace de ces observations appelle un traitement statistique adapté. C'est la raison pour laquelle cette thèse s'inscrit dans l'utilisation de l'*inversion* dans le cadre des *statistiques bayésiennes* qui sont particulièrement adaptées à notre problème.

L'approche bayésienne pour la résolution de problèmes inverses se prépare en trois étapes.

Modèle physique direct Connaissant l'objet d'intérêt et la physique des instruments de mesures, on met en place un modèle mathématique de l'observation que l'on acquiert. Évidemment, il doit être assez précis pour bien décrire et prédire les données, tout en restant suffisamment simple pour que l'inversion reste faisable.

Modèle probabiliste direct Ensuite, on décrit les phénomènes physiques et les incertitudes associées par des probabilités. Cette « traduction » du modèle physique en un modèle probabiliste propage aussi certaines propriétés, comme notamment des indépendances dans les modèles hiérarchiques.

Problème inverse Cette étape a pour but d'estimer la quantité d'intérêt étant donné l'observation **et** le modèle décrit dans les étapes précédentes.

1. qui est d'un point de vue moderne difficile à lire et manque de rigueur mathématique

Tab. 2.1 Tableau récapitulatif des différences entre les aspects fréquentiste et bayésien (complété à partir de [24])

Qu'est-ce que ...	**pour un fréquentiste**	**pour un bayésien**
probabilité	une fréquence d'apparition	une confiance (« belief »)
objectivité	dépendant des données uniquement	dépendant des données et de l'*a priori*
calcul	souvent faisable	faisable, mais parfois compliqué
flexibilité	quelques applications nécessistent simplifications	pas de limitations intrinsèques
aléatoire	pour chaque paramètre, il existe une valeur vraie fixée mais inconnue, l'observation est un événement parmi beaucoup d'événements de données possible	chaque paramètre est une variable aléatoire de valeur inconnue, les données sont connues
test d'hypothèse	rejet de l'hypothèse nulle avec erreur fixée	choix de l'hypothèse la plus probable
choix du modèle	parfois possible	direct (choix du modèle le plus probable)
région de confiance	Une région aléatoire C a une probabilité donnée de contenir le paramètre θ.	θ a une probabilité donnée d'appartenir à une région fixée C.

Au long de ce manuscrit, nous allons décrire ces trois étapes pour notre application.

Pour l'anecdote, Sir Arthur Conan Doyle exprime la différence entre problème direct et problème inverse à travers de son personnage célèbre Sherlock Holmes :

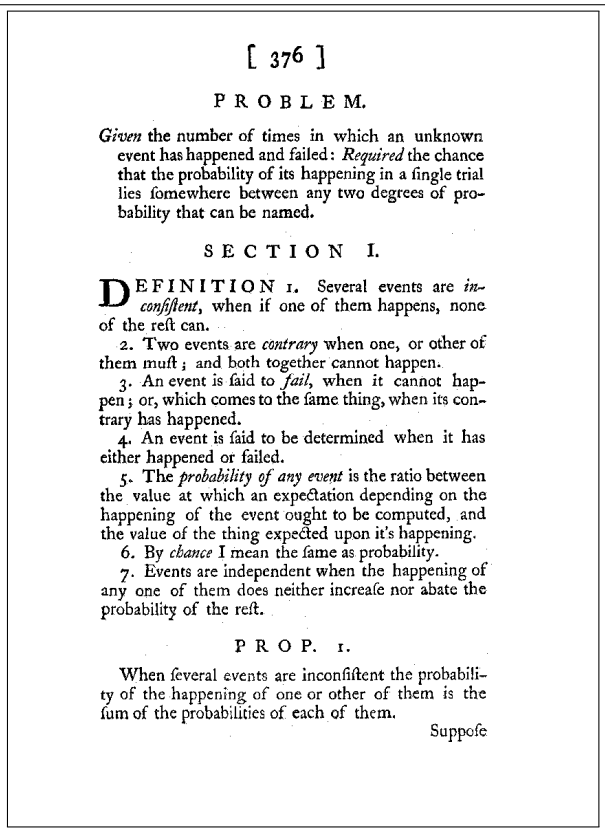
> *« Most people, if you describe a train of events to them, will tell you what the result will be. They can put those events together in their minds, and argue from them that something will come to pass. There are few people, however, who, if you told them a result, would be able to evolve from their own inner consciousness what the steps were which led up to that result. This power is what I mean when I talk of reasoning backward, or analytically. »*
>
> *Sherlock Holmes dans* A Study in Scarlet *[26]*

Notons qu'une estimation – quelle que soit la méthode – n'est jamais parfaite, même à partir de mesures directes.

> *« By the principle of inverse probability we shall be able then to proceed from the observations to estimates of the true values of the parameters, which, however, will not be exact determinations, but will have ranges of uncertainty corresponding to the fact that the individual random errors in the observations are not definitely known. »*
>
> *[19, page 73]*

Ainsi, une estimation doit toujours être accompagnée d'une marge d'erreur ou d'un *intervalle de crédibilité*, voir Tab. 2.1. La résolution du problème inverse se basant sur la loi de probabilité *a posteriori* du paramètre d'intérêt, nous avons un accès immédiat à l'intervalle de crédibilité du paramètre.

Fig. 2.1 Extrait de l'essai de Th. Bayes [27, page 7 du document numérique]



[376]

PROBLEM.

Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a ſingle trial lies ſomewhere between any two degrees of probability that can be named.

SECTION I.

DEFINITION 1. Several events are *inconſiſtent*, when if one of them happens, none of the reſt can.

2. Two events are *contrary* when one, or other of them muſt; and both together cannot happen.

3. An event is ſaid to *fail*, when it cannot happen; or, which comes to the ſame thing, when its contrary has happened.

4. An event is ſaid to be determined when it has either happened or failed.

5. The *probability of any event* is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon it's happening.

6. By *chance* I mean the ſame as probability.

7. Events are independent when the happening of any one of them does neither increaſe nor abate the probability of the reſt.

PROP. 1.

When ſeveral events are inconſiſtent the probability of the happening of one or other of them is the ſum of the probabilities of each of them.

Suppoſe

2.2 Qu'est-ce que « le bayésien » ?

Le terme bayésien (utilisé pour la première fois en tant que tel dans les années 1950) réfère au *Reverend Thomas Bayes* (1702-1761)[i] qui dans son essai [27] introduit la question suivante :

> « Given *the number of times in which an unknown event has happened and failed:* Required *the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.* »

Dans des termes modernes, Thomas Bayes veut savoir à partir des observations d'un nombre de tirages de type Bernoulli (disons « succès » contre « échec »), quelle est la probabilité *a posteriori* d'avoir par exemple une sortie « succès », étant donné une loi *a priori* uniforme [102]. Ceci est noté comme la première utilisation d'un cas spécial de ce qui sera appelé plus tard *règle de Bayes*

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A)}. \tag{2.1}$$

En mots : la probabilité d'observer B sachant A est donnée par le produit de la probabilité d'observer A sachant B par la probabilité d'avoir B, divisé par la probabilité d'avoir A. (Pierre-Simon Laplace (1749-1827)[ii] a démontré une version plus générale de ce théorème, apparemment indépendamment des travaux de Bayes [28].)

Il s'agit alors d'inférer sur un paramètre (comme ci-dessous) ou sur une sortie non encore observée (comme dans la question de Thomas Bayes) en termes de « probabilité ».

2.3 Distributions a priori, distributions a posteriori

Le théorème de Bayes fait le lien entre deux probabilités conditionnelles. Nous réécrivons l'Éqn. (2.1) sous la forme suivante

$$\mathrm{p}(\theta \mid \mathbf{Y}) = \frac{\mathrm{p}(\mathbf{Y} \mid \theta)\,\mathrm{p}(\theta)}{\mathrm{p}(\mathbf{Y})}. \tag{2.2}$$

Dans cette équation, nous avons
- le paramètre inconnu : θ,
- les données observées : $\mathbf{Y}$,
- la probabilité des données sachant le paramètre, aussi appelée *distribution d'échantillonnage* et *vraisemblance du paramètre θ attachée aux données* $\mathbf{Y}$: $\mathrm{p}(\mathbf{Y} \mid \theta)$,
- la probabilité « *a priori* » du paramètre : $\mathrm{p}(\theta)$,
- la probabilité « *a posteriori* » du paramètre : $\mathrm{p}(\theta \mid \mathbf{Y})$.

En écrivant ceci, nous considérons alors le paramètre θ comme une variable aléatoire (et non comme une quantité fixe non aléatoire). Cette *probabilisation* se fait dans deux sens :
- **a priori**, c'est-à-dire avant l'acquisition des données, l'expérience, l'observation, la mesure, intégrant les connaissances *a priori* (expert, littérature, apprentissage, ...),
- **a posteriori**, c'est-à-dire en utilisant les évidences qu'apportent les données dans la connaissance de la valeur du paramètre, apportées par la distribution d'échantillonnage $\mathrm{p}(\mathbf{Y} \mid \theta)$ qui peut également être interprétée comme un terme d'attache aux données par rapport au paramètre. [2]

La vraisemblance d'un paramètre θ, *i.e.* la probabilité de la sortie sachant le paramètre *vue comme fonction de ce dernier* $L(\theta) = \mathrm{p}(\mathbf{Y} \mid \theta)$ est en lien étroit avec le modèle direct, et notamment avec la modélisation du bruit. C'est sa distribution qui détermine celle des données et donc la fonction de vraisemblance du paramètre. Ainsi, nous probabilisons le problème direct.

Notons que la fonction de vraisemblance de deux paramètres distincts n'a pas forcément la même forme alors qu'on part de la même distribution d'échantillonnage. Considérons par exemple le cas d'une distribution normale des données avec une moyenne μ et un écart-type σ : $\mathrm{p}(\mathbf{Y} \mid \mu, \sigma)$, et notons la vraisemblance du paramètre μ avec $\sigma = \sigma_0$ fixée $L_{\sigma_0}(\mu) = \mathrm{p}(\mathbf{Y} \mid \mu, \sigma_0)$ et celle de σ avec $\mu = \mu_0$ fixée $L_{\mu_0}(\sigma) = \mathrm{p}(\mathbf{Y} \mid \mu_0, \sigma)$. Alors que la distribution d'échantillonnage est la même, les fonctions de vraisemblances n'ont pas les mêmes comportements, propriétés et formes, la première prenant l'allure d'une loi normale, la deuxième d'une loi inverse-gamma.

La loi *a priori* est généralement choisie parmi les distributions pour transcrire fidèlement les connaissances disponibles sur le paramètre avant l'observation. Cependant, les propriétés statistiques rentrent aussi en compte, notamment celle de la conjugaison.

2.3.1 Conjugaison

Une propriété utile pour le choix de lois *a priori* est celle de la conjugaison de ces dernières par leurs vraisemblances. Quand elle est vérifiée, une loi *a posteriori* pour un paramètre donné a la même forme que la loi *a priori* associée. Ceci nous permettra d'avoir explicitement la loi *a posteriori* sous forme d'une loi usuelle, comme une loi normale, gamma, *et cetera*.

2. Certains auteurs parlent également de modulation de la loi *a priori* par la vraisemblance pour désigner la loi *a posteriori* [29].

Définition ⟨2.1⟩ (Lois conjuguées) *Une famille $\mathcal{F}$ de distributions de probabilité sur Θ est dite* ***conjuguée*** *par une fonction de vraisemblance $\mathrm{p}(\mathbf{Y} \mid \theta)$ si pour tout $\mathrm{p} \in \mathcal{F}$ la distribution a posteriori $\mathrm{p}(\theta \mid \mathbf{Y})$ est également élément de la famille $\mathcal{F}$. [2, Def. 3.3.1]*

Cette propriété sera notamment utilisée dans les méthodes d'échantillonnage stochastique (Sect. 2.9.1) pour les lois *a posteriori* conditionnelles.

Les références [2, Tab. 3.3.1, p. 121] et [103] ↓3 présentent des tableaux de lois *a priori* conjuguées par les vraisemblances associées.

2.4 Le dilemme de l'a priori impropre

Afin de modéliser un manque de connaissance sur le paramètre θ, le choix de la distribution *a priori* conduit souvent à une distribution *impropre*. Nous nous proposons dans cette section de définir les termes d'*a priori* non informative et d'*a priori* impropre à travers le souhait de n'injecter aucune information extérieure, puis de montrer les dangers — souvent omis ou ignorés — qui sont associés à ces lois ↓4. Nous justifions pourquoi nous évitons ces distributions dans cette thèse, à la fois dans cette section « théorique », puis ensuite quand nous parlerons de risque dans la Sect. 2.5 et d'approximation des lois *a posteriori* par échantillonnage stochastique, Sect. 2.9.1.

2.4.1 A priori non informative

Dans certaines situations, on souhaite ne pas inclure d'information extérieure dans la modélisation probabiliste d'un problème donné. Sir Harold Jeffreys[iii] l'exprime de la manière suivante :

> *« But there is one [problem] at the beginning: how can we assign the prior probability when we know nothing about the value of the parameter [...]? The answer is really clear enough when it is recognized that a probability is merely a number associated with a degree of reasonable confidence and has no purpose except to give it a formular expression. If we have no information relevant to the actual value of a parameter, the probability must be chosen so as to express the fact that we have none. »*
>
> *[19, page 118]*

On choisit alors une loi de probabilité qui représente l'ignorance et qui ajoute le moins d'information *a priori* possible, voire même aucune, dans la construction de la loi *a posteriori*. Ceci peut être traduit par la devise « laissons parler les données » afin que l'inférence ne soit pas perturbée par une information extérieure aux données.

Définition ⟨2.2⟩ *Soit $\mathrm{p}(\theta)$ la loi de probabilité* a priori *du paramètre θ de telle sorte qu'elle formalise une ignorance quant à sa valeur. Une telle distribution* a priori *est appelée* non informative.

Où rencontre-t-on des lois *a priori* non informatives ? Ce type de distributions est souvent utilisé pour modéliser des hyperparamètres. Ces derniers pilotent d'autres lois (ainsi devenues informatives) qui sont importantes dans la modélisation du problème donné (voir 2.2(c) sur la page 28 pour avoir un exemple graphique : la loi de μ est pilotée

3. Même s'il s'agit d'un document en ligne issu de l'encyclopédie participative *Wikipedia* qui n'est pas une source sure et manque parfois de clarté et de rigueur dans les notations, aucun autre document aussi concis et complet nous est connu, d'où sa citation.

4. En effet, lors de l'analyse des résultats des algorithmes développés dans cette thèse, nous avons pu remarquer des résultats aberrants non attendus quand nous utilisions des lois impropres. L'analyse de ces résultats a débouché en la considération de ce que nous présentons dans la présente section.

par les hyperparamètres μ_0 et σ_0 [5]). Malheureusement, on n'a souvent « aucune » [6] information sur ces hyperparamètres. C'est la raison pour laquelle les lois non informatives semblent s'imposer.

Une question importante est celle de la construction d'une loi non informative [31]. Laplace et Bayes répondraient en disant que la meilleure façon de décrire un manque d'information est d'utiliser une loi uniforme [32]. Ainsi, toute valeur possible pour ce paramètre a la même probabilité et s'oppose à ce que Laplace appela *principe de cause suffisante* [33]. Cependant, si on n'a pas d'information sur θ, on n'en a pas non plus sur ses transformations. Ainsi, la loi uniforme pour θ ne traduit pas une information équivalente par rapport à la loi uniforme pour $1/\theta$, $\exp(\theta)$, $\log\theta$, *etc.*

Pour tenir compte de la transformation, soient Ω et Θ deux espaces de probabilité, ω et θ des variables aléatoires et $g : \Omega \to \Theta$ une fonction inversible. Choisissons ensuite une loi *a priori* $\mu(\omega)$ sur Ω. Avec la transformation $\theta = g(\omega)$, on définit également l'*a priori* $\pi = \mu \circ g$ sur Θ car pour un $A \subseteq \Theta$, on a $\pi(A) = \mu(g^{-1}(A))$. La loi uniforme par exemple est invariante par rapport aux permutations à l'intérieur d'un ensemble fini et aux translations affines dans le cas continu. L'étude de l'invariance par rapport aux transformations, dont les premiers travaux ont été entrepris par [19, 34, 35], nous renseigne sur le choix de l'*a priori*.

Principe de Jeffreys

Pour Jeffreys [34], [19, Sect. 3.10], l'invariance par rapport aux changements de variables ou d'échelle est le fondement du choix d'une loi *a priori*. L'article [35] étend les travaux de Jeffreys et propose un résumé de lois invariantes. Le principe de Jeffreys mène finalement à la définition de lois non informatives [3, Sect. 2.9] que l'on appelle aussi *a priori de Jeffreys*. Il est basée sur l'information de Fisher[iv] [36, Sect. 2.2] qui est un indicateur de l'information moyenne sur le paramètre θ apportée par l'observation $\mathbf{Y}$:

$$J(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial \log \mathrm{p}(\mathbf{Y} \mid \theta)}{\partial\theta}\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2 \log \mathrm{p}(\mathbf{Y} \mid \theta)}{\partial\theta^2}\right]. \tag{2.3}$$

Définition ⟨2.3⟩ *L'a priori de Jeffreys est définie par*

$$\mathrm{p}(\theta) \propto [J(\theta)]^{1/2}. \tag{2.4}$$

L'information apportée par cette loi est invariante par rapport aux changements de variables : l'information de Fisher du paramètre $\phi = h(\theta)$ par rapport à la fonction de vraisemblance $\tilde{\mathrm{p}}(\mathbf{Y} \mid \phi)$ est calculée en utilisant les définitions ci-dessus et les règles de dérivation : $J(\phi) = J(h(\theta))h'(\theta)^2$ et $\mathrm{p}(\phi) \propto \left[J(h(\theta))h'(\theta)^2\right]^{1/2} = \mathrm{p}(h(\theta))\,|h'(\theta)|$. On reconnaît dans la dernière expression la formule de transformation de variables ce qui valide la propriété souhaitée. En résumé, si θ suit la loi *a priori* de Jeffreys $\mathrm{p}(\theta) \propto J(\theta)^{1/2}$, alors une reparamétrisation bijective $\phi = h(\theta)$ mène à la loi *a priori* $\mathrm{p}(\phi) \propto J(\phi)^{1/2}$ pour ϕ qui est également de Jeffreys. La loi *a priori* ne dépendant plus de la paramétrisation de la variable aléatoire, la loi *a posteriori* en est indépendante aussi [37] car l'expression de la loi *a posteriori* à partir de $\mathrm{p}(\theta)$ implique celle déduite à partir de $\mathrm{p}(\phi)$ par transformation.

Notons aussi que la loi *a priori* déterminée par le principe de Jeffreys est dépendante de fonction la vraisemblance par l'utilisation de l'information de Fisher. Plus fort même,

5. Dans la référence [30] associée à ce diagramme, les hyperparamètres sont appris au préalable à l'aide d'une cohorte d'apprentissage ; les auteurs ne donnent cependant pas plus d'information sur cette étape.

6. Les guillemets se justifieront par la suite.

les inférences sur le paramètre θ que l'on entreprend par rapport à cette loi ne dépendent pas d'**une** vraisemblance, mais de la vraisemblance moyenne par rapport à θ puisqu'il s'agit d'une espérance sur θ.

Exemple ⟨2.4⟩ *Illustrons le principe de Jeffreys par deux exemples. Soit pour cela la vraisemblance* $\mathrm{p}(\mathbf{Y} \mid m, \gamma)$ *une loi normale mono-variée de moyenne m et de précision* γ *:*

$$\mathrm{p}(\mathbf{Y} \mid m, \gamma) = \left(\frac{\gamma}{2\pi}\right)^{1/2} \exp\left(-\frac{\gamma}{2}(m - \mathbf{Y})^2\right).$$

1. *Pour ce premier exemple, soit* $m = m_0$ *fixe. Nous cherchons donc une loi a priori pour* γ *en utilisant le principe de Jeffreys. L'information de Fisher par rapport au paramètre d'intérêt s'écrit*

$$J(\gamma) = -\mathbb{E}_\gamma\left[-\frac{1}{2}\frac{\partial^2}{\partial\gamma^2}\left(\log 2\pi - \log\gamma + \gamma(m_0 - \mathbf{Y})^2\right)\right] = \mathbb{E}_\gamma\left[\frac{1}{2}\gamma^{-2}\right] = \gamma^{-2}/2\,.$$

   En utilisant la définition de l'Éqn. (2.4), on obtient $\mathrm{p}(\gamma) \propto \gamma^{-1}$*. On note que cette loi a priori n'est pas intégrable.*

2. *Soit maintenant la précision* $\gamma = \gamma_0$ *fixe, et la moyenne m le paramètre d'intérêt. L'information de Fisher s'exprime de la manière suivante*

$$J(m) = -\mathbb{E}_m\left[-\frac{1}{2}\frac{\partial^2}{\partial m^2}\left(\log 2\pi - \log\gamma_0 + \gamma_0(m - \mathbf{Y})^2\right)\right] = \gamma_0\mathbb{E}_m\left[1\right] = \gamma_0.$$

   Ainsi, la loi a priori pour le paramètre de la moyenne issue du principe de Jeffreys est une loi uniforme sur l'espace de probabilité $\Omega_m = \mathbb{R}$, $\mathrm{p}(m) \propto 1$*. On note que cette loi a priori n'est pas intégrable.*

2.4.2 A priori impropre

Les distributions *a priori* impropres surviennent quand des probabilités sont définies non sur un ensemble fini d'événements, mais sur un ensemble infini (par exemple l'ensemble des entiers naturels $\mathbb{N}$) ou un continuum d'événements (par exemple l'ensemble des réels $\mathbb{R}$). Alors que l'axiome d'additivité des probabilités des événements peut être étendu facilement, il n'en est pas de même pour l'axiome qui demande que la somme des probabilités fasse 1 [23]. Comment définir sur l'ensemble $\mathbb{N}$ par exemple une loi uniforme se sommant à 1 si la cardinalité est infinie ? Est-ce que l'on accepte une « distribution » non intégrable (au sens de la mesure en question, *i.e.* dans l'exemple précédent la mesure de comptage sur $\mathbb{N}$) ? D'après [23, Ch. 2], il n'y a pas de théorie formelle pour gérer ces « lois de probabilités ». Ceci appelle la définition suivante [2, Ch. 1.5] :

Définition ⟨2.5⟩ *Soit* $f : \Omega_\theta \to [0, \infty[$ *une mesure telle que*

$$\int_{\Omega_\theta} f(\theta)\,\mathrm{d}\theta = +\infty, \tag{2.5}$$

i.e. f est une mesure σ-finie, mais pas une mesure de probabilité. Si f est associé à une variable aléatoire θ en tant que loi a priori, elle est appelée loi *a priori* impropre *(ou loi a priori généralisée).*

Comme il ne s'agit plus d'une densité de probabilité avec intégrale 1, la règle de Bayes ne tient plus. Cependant, on *définit* la loi *a posteriori* $\mathrm{p}(\theta \mid \mathbf{Y})$ comme rapport entre $\mathrm{p}(\mathbf{Y} \mid \theta)\,\mathrm{p}(\theta)$ et $\mathrm{p}(\mathbf{Y})$.

Il est important de noter que la construction d'une loi non informative débouche souvent en une loi impropre, comme nous avons vu dans l'exemple précédent : l'intégrale de la loi non informative issue du principe de Jeffreys pour le paramètre d'inverse variance du bruit de mesure, $\mathrm{p}(\gamma) \propto \gamma^{-1}$, n'est pas finie, et il en est de même pour le paramètre de la moyenne modélisée *a priori* par $\mathrm{p}(m) \propto 1$. Parmi les conséquences que cela peut avoir [32, Sect. 4.2], on compte notamment le caractère impropre de la loi *a posteriori* totale. Si cela est le cas, la base à partir de laquelle nous souhaitons faire l'inférence sur un paramètre n'est plus une distribution, et donc l'inférence peut ne pas être valide.

2.4.3 Mises en garde

L'utilisation d'une loi *a priori* impropre a des conséquences importantes sur la loi jointe des paramètres. En effet, considérons la situation suivante : le paramètre θ est distribué sous la loi $\mathrm{p}(\theta)$ qui est une loi impropre, et la vraisemblance $\mathrm{p}(\mathbf{Y} \mid \theta)$ est intégrable en $\mathbf{Y}$ et donc propre. La loi jointe s'obtient par la multiplication des deux : $\mathrm{p}(\theta, \mathbf{Y}) = \mathrm{p}(\theta)\,\mathrm{p}(\mathbf{Y} \mid \theta)$. Vérifions maintenant l'intégrabilité de la loi jointe en bénéficiant des propriétés de marginalisation :

$$\int_{\Omega_\theta \times \Omega_\mathbf{Y}} \mathrm{p}(\theta, \mathbf{Y})\,\mathrm{d}\mathbf{Y}\,\mathrm{d}\theta = \int_{\Omega_\theta} \mathrm{p}(\theta)\,\mathrm{d}\theta = \infty.$$

En choisissant une loi *a priori* impropre, la loi jointe devient elle-même impropre !

Notons qu'en marginalisant d'abord le paramètre θ, la vérification de l'intégrabilité serait réduite à celle de $\mathrm{p}(\mathbf{Y})$ qui peut être une loi propre. Pourquoi cette contradiction ? Les conditions du théorème de Fubini permettant de changer l'ordre d'intégration ne sont pas toutes satisfaites, notamment l'intégrabilité des fonctions.

La référence [32] note dans sa section 4.2.5 que la caractérisation de lois *a priori* impropres résultant dans des lois *a posteriori* propres est toujours un problème ouvert. L'utilisation de lois impropres a également un impact sur le risque bayésien que l'on minimise dans le choix d'un estimateur, voir ci-dessous dans la Sect. 2.5.

2.4.4 Bilan

Cette section introduit les lois non informatives comme possible choix d'*a priori* pour un paramètre. Soulignons deux interprétations de ce choix. Premièrement, il s'agit d'une représentation formelle de l'ignorance par rapport à ce paramètre. Deuxièmement, il n'existe pas une loi unique qui traduit ce fait ; on la choisit plutôt par convenance, par rapport à l'utilisateur, au problème, aux calculs à expliciter, à la difficulté d'établir des lois subjectives informatives appropriées, *etc.*

L'utilisation des lois *a priori* impropres que nous avons introduites dans cette section semble dangereuse. En effet, la loi jointe résultante n'est plus une densité dans le sens où elle n'est plus intégrable alors que dans certains cas, la loi *a posteriori* obtenue l'est [38, Sect. 4.2.3]. *A contrario*, on peut avoir des lois *a posteriori* conditionnelles qui sont toutes propres, mais la loi *a posteriori* totale ne l'est pas [39]. Le caractère propre de la loi *a posteriori* est donc à vérifier pour chaque nouveau problème. Ce calcul n'est cependant pas toujours faisable, et on s'adonne à des inférences à partir d'une fonction — appelée faussement loi *a posteriori*— ce qui peut mener à des résultats aberrants.

Pour pallier ce défaut, nous proposons « simplement » de ne pas utiliser les lois impropres. En effet, même si on pense ne pas avoir d'information sur la valeur d'un paramètre, on peut dans la plupart des problèmes auxquels on se confronte au moins limiter la plage des valeurs à une région finie. Dans le cas de cette thèse par exemple, certains paramètres sont strictement positifs, ou ne dépasseront pas des valeurs limites. Cette information, aussi petite soit-elle, peut être injectée dans le modèle probabiliste. Dans des

cas où la loi impropre f que l'on aurait envie de choisir peut être exprimée comme limite d'une suite de fonctions intégrables $(f_n)_n$, *i.e.* de lois propres, il suffit de ne pas aller jusqu'à la limite mais de s'arrêter à un n' pour lequel la fonction $f_{n'}$ est suffisamment non informative pour ne pas créer de biais ou de déformation. On attribue l'adjectif *vaguement informatif* à ces lois. C'est en effet ce type de lois que nous choisissons dans cette thèse lorsque nous voulons exprimer un « manque d'information » sur un paramètre.

2.5 Estimateurs ponctuels

Dans l'inférence bayésienne, on peut préférer donner des propriétés de la loi *a posteriori*, comme la moyenne, le mode, la variance, *et cetera*. Le choix de l'estimateur retranscrivant une propriété ne se fait pas *ad hoc* mais est réfléchi en fonction de l'optimisation de la fonction mesurant l'inadéquation d'une estimation.

Le point de départ du choix d'un estimateur est la fonction de coût.

Définition ⟨2.6⟩ *Soit Ψ l'ensemble de tous les estimateurs possibles $\psi : \Omega_{\mathbf{Y}} \to \Omega_\theta$ associant à une observation $\mathbf{Y} \in \Omega_{\mathbf{Y}}$ l'estimation $\psi(\mathbf{Y}) = \hat{\theta} \in \Omega_\theta$ par rapport à la quantité vraie θ. La fonction $\mathrm{L} : \Omega_\theta \times \Omega_\theta \to [0, \infty[$, associant au couple « vérité/estimation » $(\theta, \psi(\mathbf{Y}))$ une mesure d'erreur $\mathrm{L}(\theta, \psi(\mathbf{Y}))$ positive ou nulle est appelée* fonction de coût.

Introduisons à travers les définitions suivantes deux quantités importantes pour la construction d'estimateurs optimaux.

Définition ⟨2.7⟩

1. *Soit $\mathrm{L} : \Omega_\theta \times \Omega_\theta \to [0, \infty[$ une fonction de coût, et soit $\mathrm{p}(\theta \mid \mathbf{Y})$ la distribution a posteriori du paramètre θ. Le* coût *a posteriori* $\mathrm{L}_{\Theta|\mathbf{Y}} : \Psi \to [0, \infty[$ *est défini par*

$$\mathrm{L}_{\Theta|\mathbf{Y}}(\psi) = \mathbb{E}_{\Theta|\mathbf{Y}}\left(\mathrm{L}(\Theta, \psi(\mathbf{Y}))\right) = \int_{\Omega_\theta} \mathrm{L}(\theta, \psi(\mathbf{Y}))\, \mathrm{p}(\theta \mid \mathbf{Y})\, \mathrm{d}\theta. \tag{2.6}$$

2. *Étant donnée une distribution a priori $\mathrm{p}(\theta)$ pour θ et une fonction de vraisemblance $\mathrm{p}(\mathbf{Y} \mid \theta)$, le* risque moyen $\mathrm{R} : \Psi \to [0, \infty[$ *est l'espérance du coût a posteriori sous la loi a priori pour θ,*

$$\begin{aligned} \mathrm{R}(\psi) = \mathbb{E}_\Theta\left[\mathbb{E}_{\mathbf{Y}|\Theta}\left(\mathrm{L}(\Theta, \psi(\mathbf{Y}))\right)\right] &= \int_{\Omega_\theta} \int_{\Omega_{\mathbf{Y}}} \mathrm{L}(\theta, \psi(\mathbf{Y}))\, \mathrm{p}(\mathbf{Y} \mid \theta)\, \mathrm{d}\mathbf{Y}\, \mathrm{p}(\theta)\, \mathrm{d}\theta \\ &= \int_{\Omega_{\theta,\mathbf{Y}}} \mathrm{L}(\theta, \psi(\mathbf{Y}))\, \mathrm{p}(\theta, \mathbf{Y})\, \mathrm{d}(\theta, \mathbf{Y}) \end{aligned} \tag{2.7}$$

Le théorème 2.3.2 dans [2, p. 63] montre (par application du théorème de Fubini) que ces deux définitions sont équivalentes dans le sens où les décisions finales sont les mêmes.

Définissons maintenant ce que nous entendons par « estimateur optimal » :

Définition ⟨2.8⟩ *L'*estimateur bayésien *(ou optimal) est celui qui minimise le* risque moyen,

$$\psi_{bay}(\mathbf{Y}) = \underset{\psi \in \Psi}{\operatorname{argmin}} \int_{\Omega_{\theta,\mathbf{Y}}} \mathrm{L}(\theta, \psi(\mathbf{Y}))\, \mathrm{p}(\theta, \mathbf{Y})\, \mathrm{d}(\theta, \mathbf{Y}) = \underset{\psi \in \Psi}{\operatorname{argmin}}\, \mathbb{E}_{\Theta,\mathbf{Y}}\left[\mathrm{L}(\theta, \psi(\mathbf{Y})\right]. \tag{2.8}$$

Nous introduisons dans ce qui suit deux estimateurs fréquemment utilisés dans la littérature bayésienne : le maximum et la moyenne *a posteriori*, leur construction étant détaillée dans [2].

2.5.1 Maximum A Posteriori

Sans fonction de coût explicite, un choix entre des valeurs *a posteriori* n'est pas possible [2, Ch. 4]. Si aucune fonction de coût ne peut être avancée par l'utilisateur, on choisit par défaut la fonction de coût 0–1 qui associe un coût nul à une estimation correcte et un coût unitaire à toute autre valeur.

Proposition ⟨2.9⟩ *Soit $\theta^\star \in \Omega_\theta$ un paramètre à estimer, et soit $\Psi = \{\psi : \Omega_{\mathbf{Y}} \to \Omega_\theta\}$ l'ensemble des estimateurs. L'estimateur attaché au coût* 0–1 *est l'estimateur* Maximum A Posteriori *(*MAP*) :*

$$\psi_{0\text{–}1}(\mathbf{Y}) = \hat{\theta}_{\mathrm{MAP}} = \operatorname*{argmax}_{\theta\in\Theta} \mathrm{p}(\theta \mid \mathbf{Y}). \tag{2.9}$$

Preuve
Écrivons le coût *a posteriori* pour la fonction 0–1 et minimisons-le :

$$\begin{aligned}
\operatorname*{argmin}_{\psi\in\Psi} \int_{\Omega_\theta} \mathrm{L}(\theta^\star, \psi(\theta'))\, \mathrm{p}(\theta' \mid \mathbf{Y})\, \mathrm{d}\theta' &= \operatorname*{argmin}_{\hat{\theta}\in\Omega_\theta} \int \mathrm{p}(\theta' \mid \mathbf{Y})\, \mathrm{d}\theta' - \mathrm{p}(\hat{\theta} \mid \mathbf{Y}) \\
&= \operatorname*{argmin}_{\hat{\theta}\in\Omega_\theta} 1 - \mathrm{p}(\hat{\theta} \mid \mathbf{Y}) \\
&= \operatorname*{argmax}_{\hat{\theta}\in\Omega_\theta} \mathrm{p}(\hat{\theta} \mid \mathbf{Y}).
\end{aligned}$$

La première ligne tient parce que nous explicitons le coût pour toutes les valeurs de $\theta' \in \Omega_\theta$, sachant qu'il existe $\psi(\mathbf{Y}) = \hat{\theta} \in \Omega_\theta$ tel que $\mathrm{L}(\theta^\star, \hat{\theta}) = 0$ car l'espace de l'estimation correspond à celui du paramètre. Le coût *a posteriori* est finalement minimisé en maximisant la loi *a posteriori* ce qui valide la proposition. ■

2.5.2 Espérance A Posteriori

La fonction de coût la plus fréquemment utilisée est le coût quadratique $\mathrm{L}(\theta, \psi) = (\theta - \psi(\mathbf{Y}))^2$ pour le paramètre θ et son estimé $\psi(\mathbf{Y})$. Dans la proposition suivante, nous allons déduire l'estimateur bayésien optimal attaché à ce coût.

Proposition ⟨2.10⟩ *Parmi tous les estimateurs possibles — qu'ils soient bayésiens, fréquentistes, aléatoires, ou d'une autre nature — l'estimateur* Espérance A Prosteriori *(*EAP*, aussi appelé* moyenne *a posteriori ou, en anglais, Posterior Mean), noté $\psi_2(\mathbf{Y})$, minimise l'erreur quadratique moyenne [2, Sect. 2.5], voir aussi Sect. 2.8.*

Il est défini par l'intégration du paramètre sous la mesure $\mathrm{p}(\theta \mid \mathbf{Y})\, \mathrm{d}\theta$ *:*

$$\psi_2(\mathbf{Y}) = \hat{\theta}_{\mathrm{EAP}} = \mathbb{E}_{\Theta|\mathbf{Y}}(\theta) = \int_{\Omega_\theta} \theta\, \mathrm{p}(\theta \mid \mathbf{Y})\, \mathrm{d}\theta. \tag{2.10}$$

Preuve
La démonstration de ce résultat est une simple succession de définitions. Nous exprimons d'abord l'erreur quadratique moyenne :

$$\begin{aligned}
\mathbb{E}_{\Theta,\mathbf{Y}}\left[(\theta - \psi(\mathbf{Y}))^2\right] &= \mathbb{E}_{\mathbf{Y}}\left[\mathbb{E}_{\Theta|\mathbf{Y}}\left[(\theta - \psi(\mathbf{Y}))^2\right]\right] \\
&= \mathbb{E}_{\mathbf{Y}}\left[\mathbb{E}_{\Theta|\mathbf{Y}}\left[\theta^2\right] - 2\psi(\mathbf{Y})\mathbb{E}_{\Theta|\mathbf{Y}}[\theta] + \psi(\mathbf{Y})^2\right].
\end{aligned}$$

Cette expression est minimisée pour $\psi(\mathbf{Y}) = \psi_2(\mathbf{Y}) = \mathbb{E}_{\Theta|\mathbf{Y}}[\theta]$. ■

Remarque ⟨2.11⟩ *L'estimateur* EAP *est le premier moment de la loi a posteriori. Les autres moments a posteriori d'ordre p (et ainsi, en les combinant, la variance, l'asymétrie, le kurtosis, …) se calculent de la même manière :* $\mathbb{E}(\theta^p) = \int_{\Omega_\theta} \theta^p\, \mathrm{p}(\theta \mid \mathbf{Y})\, \mathrm{d}\theta$.

Remarque ⟨2.12⟩ (Loi impropre et risque) *L'utilisation d'une loi impropre fait que le risque bayésien est infini, quel que soit l'estimateur ψ. Il n'existe donc pas d'estimateur qui minimise le risque bayésien et qui pourrait être optimal dans ce sens. Pour ces cas, on peut définir [40, Sect. 4.2][2, Sect. 2.3] l'*estimateur bayésien généralisé *qui minimise pour chaque* $\mathbf{Y}$ *l'espérance de la fonction de coût sous la loi* a posteriori $\mathrm{p}(\theta \mid \mathbf{Y})$. *L'utilisation de lois* a priori *intégrables évite la nécessité de cette définition. Le lecteur intéressé trouvera cependant dans les références citées ci-dessus une description plus détaillée.*

2.5.3 Intervalle de crédibilité

Comme nous l'avons déjà mentionné, entre autre en citant Jeffreys, l'estimation $\hat{\theta}$ du paramètre θ ne peut être parfaite et doit être accompagnée d'indications quant aux incertitudes. On peut alors choisir des *régions de confiance* $C \in \Omega_\theta$ dans lesquelles le paramètre devrait se trouver avec une forte probabilité. Le paradigme bayésien propose la notion de *crédibilité* [2, Sect. 5.5.1] qui donne une probabilité au paramètre θ d'appartenir à une région (fixée) C :

Définition ⟨2.13⟩ *Un ensemble C est appelé α-*crédible *si* $\Pr(\theta \in C \mid \mathbf{Y}) \geq 1 - \alpha$, *i.e. si la probabilité* a posteriori *d'appartenance de θ vaut au moins $1 - \alpha$.*

*Parmi tous les ensembles α-crédibles, celui de volume minimal et donc le plus restreint est appelé α-*crédible à plus forte densité *a posteriori*.

La détermination analytique de la région α-crédible à plus forte densité peut être difficile dans des cas compliqués et à grande dimension ; numériquement, on peut s'appuyer sur des échantillons de la loi *a posteriori* qui permettent de calculer la variance *a posteriori*.

Le vrai paramètre θ pour lequel on a obtenu une estimation $\hat{\theta}$ n'est cependant pas forcément élément de l'intervalle α-crédible associée à cette estimation. En effet, d'après la définition, θ a une probabilité d'une valeur plus petite que α de ne pas appartenir à cette intervalle (avec égalité pour l'intervalle α-crédible à plus forte densité *a posteriori*).

Comme cas limite, notons que l'intervalle α-crédible C est l'ensemble de définition Ω_θ si $\alpha = 0$, c'est-à-dire par souci de ne pas vouloir commettre d'erreur, on propose comme intervalle de crédibilité toutes les valeurs probables *a priori* pour θ. L'intervalle à plus forte densité *a posteriori* est restreint à l'ensemble vide pour $\alpha = 1$ [7].

La détermination de la région de crédibilité dépend, comme l'estimation, de la modélisation (physique et probabiliste) du problème. Un fort décalage entre vérité et estimation avec ses incertitudes peut alors nous renseigner sur une éventuelle inadéquation des modèles ou des choix lors de la sélection des lois *a priori* du paramètre même si l'estimation semble « stable » ou « converger ».

2.6 Test d'hypothèse, choix de modèle

Nous introduisons dans cette section une mesure d'évidence bayésienne. En effet, le paradigme bayésien nous permet de « probabiliser » des objets comme les hypothèses, les modèles, *et cetera*. Pour cela, il suffit de considérer une hypothèse, un modèle, une classe comme variable aléatoire et de lui attribuer une probabilité *a priori* et de la vrai-

7. car $\Pr(\theta \in C \mid \mathbf{Y}) \geq 0$ est vrai pour tout singleton $\{\theta_0\}$, $\forall \theta_0 \in \Omega_\theta$, mais $\Pr(\theta \in C \mid \mathbf{Y}) = 0$ — caractéristique pour l'intervalle à plus forte densité — correspond à l'intersection de tous ces ensembles qui est vide

semblance. Nous pouvons donc décrire explicitement la probabilité de chacune des hypothèses, classes et modèles [8] et procéder à une estimation.

Cette section introduit deux méthodes courante en statistiques bayésiennes : le facteur de Bayes et la prise de décision *via* une fonction de coût. Ensuite, nous détaillons un cas spécial de ce dernier point.

2.6.1 Facteur de Bayes

Le facteur de Bayes, introduit dans [19], propose de comparer les rapports des probabilités *a posteriori* sur *a priori* de deux hypothèses que nous noterons H_0 et H_1, respectivement appelées *hypothèse nulle* et *hypothèse alternative* dans la littérature « fréquentiste ». En fonction du résultat de cette comparaison, on choisira l'hypothèse la plus pertinente.

On note π_0 la probabilité *a priori* de l'hypothèse H_0 et π_1 celle de l'hypothèse H_1. Le rapport des probabilités *a priori*

$$\frac{\pi_0}{\pi_1} = \frac{\Pr(H_0)}{\Pr(H_1)} \tag{2.11}$$

décrit combien de fois l'hypothèse nulle H_0 est *a priori* plus probable que l'hypothèse H_1.

Considérons maintenant p_0 la probabilité *a posteriori* de l'hypothèse H_0 et p_1 celle de l'hypothèse H_1, obtenu à l'aide des données et grâce à la règle de Bayes. Le rapport des probabilités *a posteriori*

$$\frac{p_0}{p_1} = \frac{\Pr(H_0 \mid \mathbf{Y})}{\Pr(H_1 \mid \mathbf{Y})} = \frac{\Pr(H_0)\,\mathrm{p}(\mathbf{Y} \mid H_0)}{\Pr(H_1)\,\mathrm{p}(\mathbf{Y} \mid H_1)} = \frac{\pi_0}{\pi_1}\frac{\mathrm{p}(\mathbf{Y} \mid H_0)}{\mathrm{p}(\mathbf{Y} \mid H_1)} \tag{2.12}$$

exprime le degré de confiance que l'on peut avoir dans l'hypothèse H_0 au vu des *a posteriori*.

Définition ⟨2.14⟩ *Dans la situation décrite et avec les notations introduite ci-dessus, le quotient des rapports* a posteriori *et* a priori*, exprimé par*

$$\mathrm{BF}_{01} = \frac{p_0/p_1}{\pi_0/\pi_1} = \frac{\mathrm{p}(\mathbf{Y} \mid H_0)}{\mathrm{p}(\mathbf{Y} \mid H_1)} \tag{2.13}$$

est appelé facteur de Bayes. *Il est une mesure d'évidence fournie par les données en faveur de l'hypothèse H_0 ([20, Sect. 8.1], [2, Sect. 5.2.2] et [19, Ch. 5]).*

Remarque ⟨2.15⟩ $\mathrm{BF}_{10} = (\mathrm{BF}_{01})^{-1}$ *est la mesure d'évidence en faveur de l'hypothèse H_1.*

De plus, on voit aisément que pour trois hypothèses H_0, H_1 et H_2, la facteur de Bayes BF_{02} *en faveur de l'hypothèse H_0 par rapport à H_2 peut être déduit des facteurs de Bayes entre* BF_{01} *et* BF_{12} *:* $\mathrm{BF}_{02} = \mathrm{BF}_{01}/\mathrm{BF}_{12}$.

On peut se donner une échelle absolue du degré de certitude en faveur de l'hypothèse H_0 qui utilise le logarithme décimal. Une échelle similaire plus rudimentaire a été proposée par Jeffreys [19, § 1.6]. La liste de l'interprétation du facteur de Bayes, voir Tab. 2.2, est non exhaustive, les valeurs étant choisies arbitrairement et ayant un caractère quelque peu philosophique ; par contre, elle dit toujours la même chose : plus le facteur de Bayes est grand, plus les données renforcent l'hypothèse H_0 ; plus l'inverse du facteur de Bayes est grand, plus l'hypothèse H_1 est favorisée.

8. Notons que dans les approches « classiques », le seul but d'un test est le **rejet** (ou non) de l'hypothèse H_0, avec une erreur associée à la décision. L'utilisation du bayésien va plus loin : on **choisit** une hypothèse selon la probabilité qui lui est associée. Il en est de même pour la classification et la sélection de modèles.

Tab. 2.2 Échelle absolue d'évaluation du degré de certitude de l'hypothèse H_0.

$\left\{ \begin{smallmatrix}(-1)\\(+1)\end{smallmatrix} \right\} \cdot \log_{10}(\mathrm{BF}) \in \ldots$	la certitude que H_0 est $\left\{ \begin{smallmatrix}\text{fausse}\\\text{vraie}\end{smallmatrix} \right\}$ est ...
$[0, 0.5]$	faible
$[0.5, 1]$	substantielle
$[1, 2]$	forte
$[2, \infty[$	décisive

Exemple ⟨2.16⟩ *Nous souhaitons tester les hypothèses*

$$H_0 : \ \theta \in \Theta_0, \qquad H_1 : \ \theta \in \Theta_1$$

où Θ_0 et Θ_1 forment une partition de l'espace de probabilité de θ : $\Theta_0 \cup \Theta_1 = \Theta$. La loi des données $\mathbf{Y}$ conditionnellement au paramètre θ est donnée par $\mathrm{p}(\mathbf{Y} \mid \theta)$.

Si $\mathrm{p}(\theta)$ désigne la loi a priori *de θ, nous pouvons écrire le rapport des lois* a priori *des hypothèses :*

$$\frac{\pi_0}{\pi_1} = \frac{\Pr(\theta \in \Theta_0)}{\Pr(\theta \in \Theta_1)} = \frac{\int_{\Theta_0} \mathrm{p}(\theta)\,\mathrm{d}\theta}{\int_{\Theta_1} \mathrm{p}(\theta)\,\mathrm{d}\theta}. \tag{2.14}$$

À l'aide des données et de la règle de Bayes, nous pouvons également calculer le rapport des lois a posteriori *donné par*

$$\frac{p_0}{p_1} = \frac{\Pr(\theta \in \Theta_0 \mid \mathbf{Y})}{\Pr(\theta \in \Theta_1 \mid \mathbf{Y})} = \frac{\int_{\Theta_0} \mathrm{p}(\theta \mid \mathbf{Y})\,\mathrm{d}\theta}{\int_{\Theta_1} \mathrm{p}(\theta \mid \mathbf{Y})\,\mathrm{d}\theta} = \frac{\int_{\Theta_0} \mathrm{p}(\theta)\,\mathrm{p}(\mathbf{Y} \mid \theta)\,\mathrm{d}\theta}{\int_{\Theta_1} \mathrm{p}(\theta)\,\mathrm{p}(\mathbf{Y} \mid \theta)\,\mathrm{d}\theta}. \tag{2.15}$$

Le facteur de Bayes BF_{01} s'écrit donc

$$\mathrm{BF}_{01} = \frac{p_0/p_1}{\pi_0/\pi_1} = \frac{\int_{\Theta_0} \mathrm{p}(\theta)\,\mathrm{p}(\mathbf{Y} \mid \theta)\,\mathrm{d}\theta}{\int_{\Theta_1} \mathrm{p}(\theta)\,\mathrm{p}(\mathbf{Y} \mid \theta)\,\mathrm{d}\theta} \cdot \frac{\int_{\Theta_1} \mathrm{p}(\theta)\,\mathrm{d}\theta}{\int_{\Theta_0} \mathrm{p}(\theta)\,\mathrm{d}\theta}.$$

Poursuivons les calculs pour un cas simple. Soient $\mathrm{p}(\theta) = \mathcal{N}(\theta; m^{\mathrm{prior}}, \gamma^{\mathrm{prior}})$ la distribution a priori *pour θ et $\mathrm{p}(\mathbf{Y} \mid \theta) = \mathcal{N}(\mathbf{Y}; \mathbf{H}\theta, \gamma_{\mathbf{Y}})$ celle des données. Cette dernière est la fonction de vraisemblance pour θ. Suite à la conjugaison des lois, la loi* a posteriori *pour θ est également normale, avec moyenne m^{post} et précision γ^{post}.*

Soient $\Theta_0 =]-\infty, a]$ et $\Theta_1 =]a, \infty]$ les intervalles des hypothèses nulle et alternative. Le calcul du facteur de Bayes peut se faire analytiquement, le résultat étant

$$\mathrm{BF}_{01} = \frac{1 + \mathrm{erf}\left(\sqrt{\frac{\gamma^{\mathrm{post}}}{2}}(a - m^{\mathrm{post}})\right)}{1 - \mathrm{erf}\left(\sqrt{\frac{\gamma^{\mathrm{post}}}{2}}(a - m^{\mathrm{post}})\right)} \cdot \frac{1 - \mathrm{erf}\left(\sqrt{\frac{\gamma^{\mathrm{prior}}}{2}}(a - m^{\mathrm{prior}})\right)}{1 + \mathrm{erf}\left(\sqrt{\frac{\gamma^{\mathrm{prior}}}{2}}(a - m^{\mathrm{prior}})\right)}.$$

erf *désigne la fonction erreur :* $\mathrm{erf}(x) = 2/\sqrt{\pi} \int_0^x \exp(-t^2)\,\mathrm{d}t$ *pour tout $x \in \mathbb{R}$.*

Si on décide de ne pas vouloir injecter d'information a priori *tout en gardant la loi normale, on fait tendre la précision* a priori *vers 0. À la limite, le facteur de Bayes se réduit au premier facteur de l'équation précédente :*

$$\mathrm{BF}_{01} = \frac{1 + \mathrm{erf}\left(\sqrt{\frac{\gamma^{\mathrm{post}}}{2}}(a - m^{\mathrm{post}})\right)}{1 - \mathrm{erf}\left(\sqrt{\frac{\gamma^{\mathrm{post}}}{2}}(a - m^{\mathrm{post}})\right)} = \frac{2}{1 - \mathrm{erf}\left(\sqrt{\frac{\gamma^{\mathrm{post}}}{2}}(a - m^{\mathrm{post}})\right)} - 1,$$

valant $BF_{01} = 1$ *pour* $a = m^{\text{post}}$.

Du point de vue « analyse », il s'agit d'une fonction $\mathrm{BF}(x) = 2/(1 - \mathrm{erf}(x)) - 1$ *qui est continue et définie sur* $\mathbb{R}_+$. *Elle admet une limite inférieure qui est atteinte quand l'argument de la fonction erreur tend vers* $-\infty$ *:*

$$\lim_{x \to -\infty} \frac{2}{1 - \mathrm{erf}(x)} - 1 = \frac{2}{1 - \underbrace{\lim_{x \to -\infty} \mathrm{erf}(x)}_{\to -1}} - 1 = 0.$$

Quand son argument tend vers $+\infty$, *l'expression diverge du fait que* $\lim_{x \to \infty} \mathrm{erf}(x) = 1$. *À l'origine, la fonction vaut* $BF(0) = 1$. *La fonction erreur étant strictement monotone et croissante, la fonction* $\mathrm{BF}(x)$ *l'est aussi.*

Remettons ces considérations analytiques dans le contexte et étudions deux cas :

1. *La moyenne de la loi* a posteriori *se trouve dans l'intervalle* $]-\infty, a]$. *L'argument* x *de la fonction erreur est positif, la valeur du facteur de Bayes* BF_{01} *est supérieur à* 1 *: l'hypothèse nulle est favorisée. Par ailleurs, plus la moyenne de la loi* a posteriori *s'éloigne de la borne* a, *plus le facteur de Bayes croît.*
2. *La moyenne de la loi* a posteriori *se trouve dans l'intervalle* $]a, +\infty]$. *L'effet contraire apparaît : l'argument* x *de la fonction erreur est négatif, et le facteur de Bayes est inférieur à* 1. *Il tend vers* 0 *quand la moyenne de la loi* a posteriori *s'éloigne de* a. *Le facteur de Bayes s'exprime donc en faveur de l'hypothèse alternative.*

Remarque ⟨2.17⟩ (Extension à plusieurs hypothèses) *L'extension du facteur de Bayes à* N *hypothèses est relativement simple : il suffit de regarder séquentiellement les hypothèses* i *et* j, $i < j < N$. *Le choix de l'hypothèse retenue se fait en comparant tous les facteurs de Bayes récoltés.*

2.6.2 Prise de décision *via* une fonction de coût

La sous-section suivante est écrite dans un point de vue « classification ». Elle ne perd cependant pas sa validité pour les tests d'hypothèses et tests de modèles.

Pour construire un classifieur $\psi : \Omega_{\mathbf{Y}} \to \mathcal{C}$ optimal, on s'intéresse notamment aux performances de classification. Introduisons une fonction de coût $\mathrm{L} : \mathcal{C} \times \mathcal{C} \to \mathbb{R}_+$ qui est une fonction qui associe au couple $(\mathrm{c}, \psi(\mathbf{Y}))$, *i.e.* vérité et estimation, une valeur non négative $\mathrm{L}(\mathrm{c}, \psi(\mathbf{Y}))$. On impose une contrainte : quelle que soit la classe c, il coûte plus cher de se tromper que d'estimer la bonne classe, soit $\mathrm{L}(\mathrm{c}, \mathrm{c}) < \mathrm{L}(\mathrm{c}, \mathrm{c}')$ pour $\mathrm{c}' \in \mathcal{C} \setminus \{\mathrm{c}\}$. Le classifieur optimal est donc celui qui minimise le risque bayésien attaché à la fonction de coût.

Remarque ⟨2.18⟩ *Si on autorisait le coût* $\mathrm{L}(\mathrm{c}, \psi(\mathbf{Y})) \equiv 0$ *quelles que soient les classes vraie et estimée, on se retrouve avec un classifieur aléatoire ou avec un classifieur constant* $\psi(\mathbf{Y}) = \mathrm{c}^0$.

On peut s'intéresser à plusieurs cas spéciaux incrémentaux :

- Une classification correcte ne coûte rien. Dans ce cas, $\mathrm{L}(\mathrm{c}, \psi(\mathbf{Y}) = \mathrm{c}) = 0$. C'est le choix usuel pour la plupart des applications.
- Une mauvaise classification dans la classe $\mathrm{c}' \in \mathcal{C}$ coûte $\ell_{\mathrm{c}'} > 0$ quelle que soit la classe vraie. (La déduction du classifieur utilisant ce coût que l'on nommera *coût pondéré* est donnée dans Ann. A.1 pour un cas binaire.)
- La fonction de coût est une fonction binaire, *i.e.* le coût d'une bonne classification est nul, celui d'une mauvaise classification, quelle que soit la classe estimée, est unitaire. C'est ce coût que l'on considéra dans la suite.

2.6.3 Cas spécial : le coût 0–1

Le coût 0–1 est une fonction de coût « classique » qui pénalise une mauvaise décision (mauvais choix d'hypothèse, de classe) d'un poids 1 et ne pénalise pas la bonne décision, donc poids 0. Ce coût fait également la liaison entre approche statistique fréquentiste et bayésien, notamment en identifiant les erreurs type I et II au risque associé à la fonction de coût, voir [2, Sect. 2.5.3] et Sect. 2.8. L'estimateur bayésien est alors

$$\psi(\mathbf{Y}) = \operatorname*{argmax}_{\mathrm{c}\in\mathcal{C}} \Pr(\mathrm{c} \mid \mathbf{Y}). \tag{2.16}$$

L'utilisation de ce coût reporte la prise de décision à l'estimateur Maximum A Posteriori pour des lois de probabilité catégorielles[9].

Exemple ⟨2.19⟩ *Considérons le cas d'école où* $\mathsf{M} = 2$. *Les probabilités se sommant à* 1, *l'estimation au sens du* MAP *est la classe qui dépasse le seuil* $1/2$.

Attention cependant à ne pas faire d'analogie : pour $\mathsf{M} \geq 2$ *ce n'est plus un simple seuil* $1/\mathsf{M}$ *qui décide du choix. (Considérer la loi* a posteriori *catégorielle avec le vecteur de probabilité* $[0,\ 2/5,\ 3/5]$ *où les deux dernières entrées dépassent* $1/3$.*)*

Remarque ⟨2.20⟩ (Facteur de Bayes et les fonctions de coût) *Considérons le cas binaire qui est la base du facteur de Bayes et considérons le cas équiprobable. On définit l'estimateur* $\psi_{\mathrm{BF}}(\mathbf{Y}) = H_0$ *si* $\mathrm{BF}_{01} > 1$ *et* $\psi_{\mathrm{BF}}(\mathbf{Y}) = H_1$ *sinon. Autrement dit, on choisit* H_0 *si le numérateur de l'expression du facteur de Bayes (Éqn. (2.13)) est plus grand que son dénominateur, donc la probabilité* a posteriori *pour* H_0 *est plus grande que celle pour* H_1. *Ceci équivaut à la définition de l'estimateur* MAP *(Éqn. (2.16)) dans le cas binaire.*

Il est également possible de pondérer le choix de l'hypothèse à partir du facteur de Bayes. Toujours avec des probabilités a priori *égales et supposant une fonction de coût*

$$\mathrm{L}(H^\star, \psi(\mathbf{Y})) = \begin{cases} 0 & \textit{si } \psi(\mathbf{Y}) = H^\star, \\ \ell_0 & \textit{si } \psi(\mathbf{Y}) = H_0 \neq H^\star, \\ \ell_1 & \textit{si } \psi(\mathbf{Y}) = H_1 \neq H^\star, \end{cases}$$

l'estimateur à base de facteur de Bayes vaut H_0 *si* $\mathrm{BF}_{01} > \ell_0/\ell_1$ *et* H_1 *sinon.*

2.7 Bayésien hiérarchique

Beaucoup de processus physiques et naturels ainsi que d'applications statistiques sont en réalité une cascade de processus présentant une certaine hiérarchie dans la structure [3, Ch. 5].

Considérons par exemple le texte de ce manuscrit. Ce texte est composé de chapitres, les chapitres de sections, les sections de paragraphes, les paragraphes de phrases, les phrases de mots, et les mots de lettres (que l'on pourrait encore séparer en voyelles et consonnes). Les lettres sont au commencement de la chaîne (on parlera du niveau supérieur), le texte final à la fin (niveau inférieur). À partir des connaissances par exemple sur la fréquence des lettres (beaucoup de *e*, *s*, *a*), les dépendances (diphtongues, certaines lettres qui peuvent être doublées, mais pas de lettres triples, ...), on peut construire un modèle hiérarchique. Ces connaissances sont évidemment dépendantes de la langue

9. La distribution catégorielle (*categorical distribution* en anglais) est une distribution discrète sur l'ensemble d'événements $\{1, \ldots, K\}$ et étend la distribution de Bernoulli à un ensemble à plus de deux événements. La fonction masse de probabilité de l'événement $k \in \{1, \ldots, K\}$ est donnée par $\mathrm{p}(X = k) = p_k$ où p_k est la probabilité d'avoir l'événement k.

considérée (la diphtongue *eau* est fréquente en français, mais pas en allemand ou en anglais, et encore plus rare dans les langues plus éloignées comme les langues slaves). La hiérarchie peut donc être étendue en rajoutant le niveau « langue » avant le niveau « lettres »

Une phrase prononcée est composée de mots prononcés, et ces mots sont composés de phonèmes qui ont certaines dépendances entre eux (p.ex. dans la langue française, on trouve rarement plus de trois consonnes à la suite, alors que dans la langue allemande ou dans le dialecte d'arabe marocain ce n'est pas aussi rare).

L'Univers est également structuré hiérarchiquement : grossièrement, l'Univers est composé de galaxies, qui sont elles-mêmes composées d'étoiles, ces dernières d'un système stellaire comportant des planètes (ou pas).

La hiérarchie est aussi présente dans la philosophie :

> *« Les lumières célestes auxquelles s'alimentent les lumières terrestres s'ordonnent entre elles selon la façon dont elles puisent les unes aux autres, de telle sorte que la plus proche de la Source première mérite davantage le nom de lumière puisqu'elle occupe le degré le plus élevé. Pour comprendre comment cette hiérarchie peut être représentée symboliquement dans le monde visible, suppose que la lumière de la lune, après être passée par la fenêtre d'une pièce, tombe sur un miroir appliqué sur un mur, se réfléchit sur le mur opposé, pour ensuite être renvoyée sur le sol, qu'ainsi elle éclaire. Tu sais donc que la lumière du sol est due à celle du mur, que celle-ci est due à celle du miroir, et que la lumière du miroir est due à celle de la lune, laquelle est due à celle du soleil, puisque c'est lui qui éclaire la lune. Ces quatre lumières sont disposées selon un ordre hiérarchique, les unes étant plus élevées et plus parfaites que les autres, et chacune ayant une certaine place et un rang qui lui est propre et qu'elle ne saurait dépasser. »* *[41]*

Finalement, nous montrons dans le Ch. 3 que les acquisitions en protéomique sont également structurées hiérarchiquement : le mélange de protéines contenant un mélange de protéines cibles contenant un mélange de peptides contenant des peptides séparés déterminant des peptides ionisés déterminant des traces donnant la mesure finale.

Remarque ⟨2.21⟩ *Précisons le terme « hiérarchique » pour ce document. Son utilisation, on le remarquera en lisant ce manuscrit, est réservée uniquement au sens* direct *de la modélisation physique et probabiliste du problème. Le terme « bayésien hiérarchique » réfère alors à l'utilisation des statistiques bayésiennes appliquées à un modèle hiérarchique direct comme il sera présenté ici.*

Pourquoi préciser cela ? Le terme peut prêter à confusion. En effet, le terme bayésien hiérarchique peut aussi être compris comme application séquentielle d'un raisonnement bayésien niveau par niveau par l'intermédiaire d'estimations ponctuelles explicites à chaque niveau. Ainsi, on remonte dans l'arbre hiérarchique construit. Nous appellerons ce procédé par le terme « inversion séquentielle ».

Nous ne nous trouvons pas dans ce cadre puisque notre inversion, grâce au cadre entièrement bayésien, a la vocation d'être une inversion « conjointe » de tous les paramètres, quel que soit le niveau hiérarchique.

2.7.1 Dépendances et indépendances conditionnelles

On peut schématiser facilement ces structures par un diagramme comportant certains niveaux de hiérarchie comme dans la Fig. 2.2 (ne pas tenir compte du code couleur). Dans le cas d'un modèle hiérarchique bayésien, les niveaux ont une certaine dépendance entre eux, plus précisément le niveau n dépend du niveau $n-1$. (Le niveau *mot* dépend du niveau *lettre* ; le niveau *galaxie* dépend du niveau *étoile* ; le niveau *peptide* dépend du niveau *protéine*.)

Définition ⟨2.22⟩ *Si les paramètres du niveau n sont résumés dans le vecteur θ_n, et si le modèle comporte N niveaux de paramètres et un niveau « sortie » (texte, phrase prononcée, l'Univers, sortie LC-MS), alors*

– *la loi jointe des paramètres de tous les niveaux $n = 1, \ldots, N$ s'écrit*

$$\mathrm{p}(\theta_1, \ldots, \theta_N) = \mathrm{p}(\theta_1) \cdot \prod_{n=2}^{N} \mathrm{p}(\theta_n \mid \theta_{n-1}), \tag{2.17}$$

– *et la sortie* $\mathbf{Y}$ *dépend uniquement du dernier niveau hiérarchique,*

$$\mathrm{p}(\mathbf{Y} \mid \theta_1, \ldots, \theta_N) = \mathrm{p}(\mathbf{Y} \mid \theta_N). \tag{2.18}$$

Les paramètres θ_{n-1} sont appelés ***hyperparamètres*** *pour le niveau n.*

Inversement, la loi $\mathrm{p}(\theta_n \mid \theta_{n-1})$ vue en fonction de θ_{n-1} est appelée ***vraisemblance hiérarchique*** *ou* ***conditionnelle****; la loi $\mathrm{p}(\mathbf{Y} \mid \theta_N)$ est appelée* ***vraisemblance au niveau des données****.*

Cette définition et les dépendances/indépendances sont schématisées dans la Fig. 2.2 par les sous-figures (a) et (b) ; un autre exemple est donné dans la sous-figure (c).

Remarque ⟨2.23⟩ (Chaîne de Markov) *Nous avons défini le modèle hiérarchique d'une manière* ***acyclique****. Ainsi, nous trouvons une ressemblance avec une chaîne de Markov à l'ordre 1 : l'élément actuel θ_n dépend de l'élément précédent θ_{n-1}, et sous ce conditionnement il est indépendant des éléments restants $\theta_{n-2}, \ldots, \theta_1$. C'est d'ailleurs en étudiant la succession des lettres du roman* Eugène Onéguine *de Pouchkine, i.e. une structure hiérarchique, que Markov a développé cette théorie stochastique qui porte son nom aujourd'hui.*

La Fig. 2.2(c) présente un modèle hiérarchique ***cyclique****, le paramètre C étant propagé à la fois directement dans les données X (en bout de hiérarchie) et indirectement en décrivant une interaction avec le paramètre H. Nous pouvons cependant rendre ce schéma acyclique en considérant le couple de paramètres (C, H) avec une loi a priori non séparable $\mathrm{p}(C, H) = \mathrm{p}(H \mid C)\,\mathrm{p}(C)$.*

Le choix d'un modèle hiérarchique bayésien peut avoir beaucoup de justifications comme le souhait de séparer informations subjectives et objectives, « dédramatiser » la non informativité des lois, modéliser la physique des expériences, *et cetera* [2, Sect. 10.2]. Deux justifications majeures sont cependant le gain en robustesse de l'analyse bayésienne par la réduction de l'arbitraire du choix des hyperparamètres, et la simplification des calculs bayésiens que nous allons démontrer dans ce qui suit.

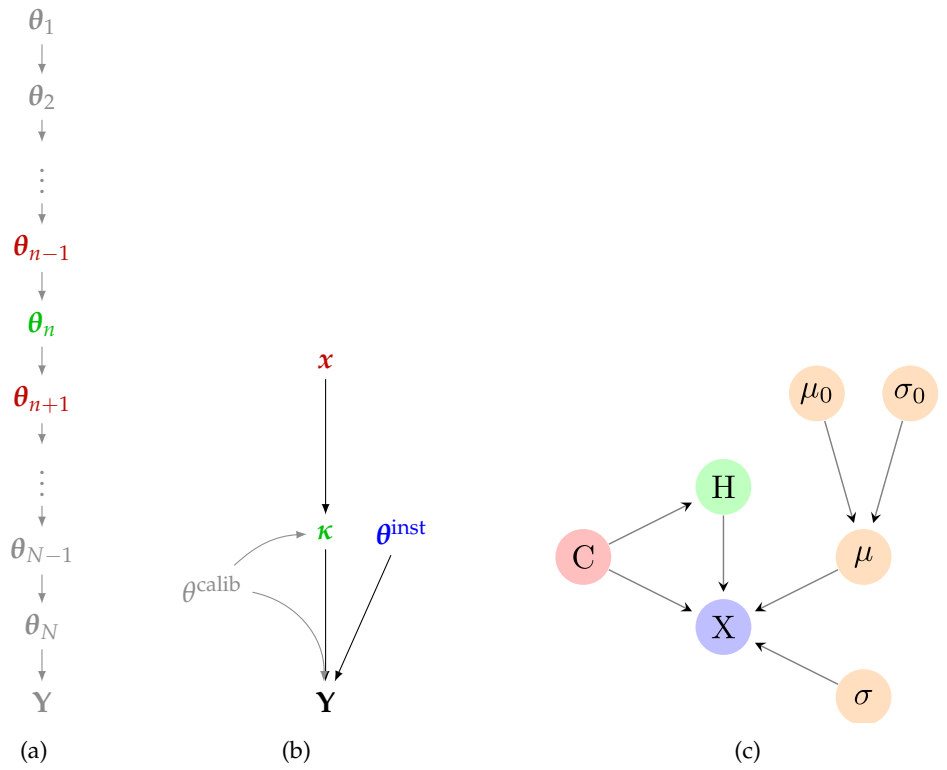
2.7.2 Lois a posteriori conditionnelles du modèle hiérarchique

Comme nous avons vu dans la Sect. 2.3, la loi *a posteriori* d'un paramètre est déterminée par sa loi *a priori* et la vraisemblance correspondant au paramètre. Dans une structure hiérarchique, ceci ne change pas : la loi *a posteriori* des paramètres au niveau hiérarchique n est proportionnelle au produit de la loi *a priori* $\mathrm{p}(\theta_n \mid \theta_{n-1})$ par la vraisemblance conditionnelle $\mathrm{p}(\theta_n \mid \theta_{n+1})$, $n = 1, \ldots, N-1$:

$$\begin{aligned}
\mathrm{p}(\theta_n \mid \mathbf{Y}, \theta_1, \ldots, \theta_{n-1}, \theta_{n+1}, \ldots, \theta_N) \propto\, & \mathrm{p}(\mathbf{Y} \mid \theta_N)\,\mathrm{p}(\theta_N \mid \theta_{N-1}) \cdots \\
& \mathrm{p}(\theta_{n+1} \mid \theta_n)\,\mathrm{p}(\theta_n \mid \theta_{n-1}) \cdots \\
& \mathrm{p}(\theta_2 \mid \theta_1)\,\mathrm{p}(\theta_1) \\
\propto\, & \mathrm{p}(\theta_{n+1} \mid \theta_n)\,\mathrm{p}(\theta_n \mid \theta_{n-1}) \\
\propto\, & \mathrm{p}(\theta_n \mid \theta_{n-1}, \theta_{n+1})
\end{aligned} \tag{2.19}$$

En identifiant le paramètre du $N+1^{\text{ème}}$ niveau avec celui des données, $\theta_{N+1} = \mathbf{Y}$, ce qui précède tient toujours et la vraisemblance conditionnelle du paramètre θ_N est la vraisemblance totale. Ainsi, $\mathrm{p}(\theta_N \mid \mathbf{Y}, \theta_1, \ldots, \theta_{N-1}) = \mathrm{p}(\theta_N \mid \mathbf{Y}, \theta_{N-1})$.

Fig. 2.2 Schéma d'une hiérarchie, réalisant les dépendances et indépendances conditionnelles. (a) Cas général vectoriel. (b) Exemple d'arbre hiérarchique sous forme de graphe acyclique orienté illustrant le modèle hiérarchique direct présenté dans le Ch. 3. (c) Exemple d'un arbre hiérarchique cyclique orienté issu de [30].



La loi *a posteriori* inclut donc uniquement les hyperparamètres directs du niveau hiérarchique qui comportent déjà les informations des niveaux supérieurs, et les paramètres résultants du niveau hiérarchique inférieur.

2.8 Mesures d'erreur

Pour quantifier l'erreur entre la vérité et l'estimation, nous introduisons ici les mesures d'erreur usuelles : biais, variance et erreur quadratique pour les estimateurs de grandeurs continues (« quantifieur ») $\psi_{\mathrm{q}} : \Omega_{\mathbf{Y}} \to \mathbb{R}$, et l'erreur de classification pour l'estimateur de grandeurs discrètes (« classifieur ») $\psi_{\mathrm{c}} : \Omega_{\mathbf{Y}} \to \mathcal{C}$. Elles servent à définir les estimateurs comme nous l'avons vu ci-dessus dans la Sect. 2.5. Dans la suite du document, elles serviront également à évaluer les performances des méthodes développées.

2.8.1 Biais

Soit $\mathbf{Y} \sim \mathrm{p}(\mathbf{Y} \mid \theta)$ une variable aléatoire, et soit θ une variable déterminée. Le **biais** est l'espérance de l'erreur d'estimation, cette dernière obtenue par l'estimateur ψ, sous la distribution d'échantillonnage des données $\mathrm{p}(\mathbf{Y} \mid \theta)$, le paramètre θ étant déterminé :

$$\mathbb{B}_{\mathbf{Y}|\Theta=\theta}(\psi, \theta) = \mathbb{E}_{\mathbf{Y}|\Theta=\theta}\left\{\theta - \psi(\mathbf{Y})\right\} = \int_{\Omega_{\mathbf{Y}}} (\theta - \psi(\mathbf{Y}))\, \mathrm{p}(\mathbf{Y} \mid \theta)\, \mathrm{d}\mathbf{Y}. \tag{2.20}$$

Le **biais moyen** s'obtient ensuite par moyennage sur toute la distribution des $\theta \sim$

$\mathrm{p}_\Theta(\theta)$ possible :

$$\mathbb{B}_{\Theta,\mathbf{Y}}(\psi) = \mathbb{E}_\Theta\left\{\mathbb{B}_{\mathbf{Y}|\Theta}(\psi,\Theta)\right\} = \mathbb{E}_{\Theta,\mathbf{Y}}\left\{\Theta - \psi(\mathbf{Y})\right\} \tag{2.21}$$

2.8.2 Variance

Dans la situation décrite ci-dessus, la **variance** de l'estimateur ψ pour un paramètre θ donné est l'espérance de l'erreur quadratique entre l'estimation $\psi(\mathbf{Y})$ et l'espérance de l'estimateur $\mathbb{E}_{\mathbf{Y}|\Theta=\theta}(\psi(\mathbf{Y}))$ [10] :

$$\mathbb{V}_{\mathbf{Y}|\theta}(\psi,\theta) = \mathbb{E}_{\mathbf{Y}|\theta}\left\{\left(\psi(\mathbf{Y}) - \mathbb{E}_{\mathbf{Y}|\theta}\left\{\psi(\mathbf{Y})\right\}\right)^2\right\} \tag{2.22}$$

qui devient en développant :

$$\mathbb{V}_{\mathbf{Y}|\theta}(\psi,\theta) = \int_{\Omega_\mathbf{Y}}\left(\psi(\mathbf{Y}) - \left[\int_{\Omega_\mathbf{Y}}\psi(\mathbf{Y})\,\mathrm{p}(\mathbf{Y}\mid\theta)\,\mathrm{d}\mathbf{Y}\right]\right)^2 \mathrm{p}(\mathbf{Y}\mid\theta)\,\mathrm{d}\mathbf{Y}.$$

La **variance moyenne** est l'espérance sous la loi de θ de cette quantité :

$$\begin{aligned}\mathbb{V}_{\Theta,\mathbf{Y}}(\psi) &= \mathbb{E}_\Theta\left\{\mathbb{V}_{\mathbf{Y}|\Theta}(\psi,\Theta)\right\}\\ &= \mathbb{E}_\Theta\left\{\mathbb{E}_{\mathbf{Y}|\Theta}\left\{\left(\psi(\mathbf{Y}) - \mathbb{E}_{\mathbf{Y}|\Theta}\left\{\psi(\mathbf{Y})\right\}\right)^2\right\}\right\}\\ &= \mathbb{E}_{\Theta,\mathbf{Y}}\left\{\left(\psi(\mathbf{Y}) - \mathbb{E}_{\mathbf{Y}|\Theta}\left\{\psi(\mathbf{Y})\right\}\right)^2\right\}.\end{aligned} \tag{2.23}$$

2.8.3 Erreur quadratique

L'**erreur quadratique** d'un estimateur ψ du paramètre à estimer θ est l'espérance du carré de l'erreur d'estimation :

$$\mathbb{EQ}_{\mathbf{Y}|\theta}(\psi,\theta) = \mathbb{E}_{\mathbf{Y}|\theta}\left\{\left(\theta - \psi(\mathbf{Y})\right)^2\right\}. \tag{2.24}$$

En introduisant le terme nul $\mathbb{E}_{\mathbf{Y}|\theta}\left\{\psi(\mathbf{Y})\right\} - \mathbb{E}_{\mathbf{Y}|\theta}\left\{\psi(\mathbf{Y})\right\}$ dans l'argument, on peut relier l'expression de l'erreur quadratique à la somme du carré du biais et la variance (voir Ann. A.2) :

$$\mathbb{EQ}_{\mathbf{Y}|\theta}(\psi,\theta) = \mathbb{B}_{\mathbf{Y}|\theta}(\psi,\theta)^2 + \mathbb{V}_{\mathbf{Y}|\theta}(\psi,\theta). \tag{2.25}$$

L'**erreur quadratique moyenne** se calcule en prenant l'espérance sous la loi de θ de l'erreur quadratique sous la loi de $\mathbf{Y}\mid\theta$:

$$\mathbb{EQ}_{\Theta,\mathbf{Y}}(\psi) = \mathbb{E}_\Theta\left\{\mathbb{EQ}_{\mathbf{Y}|\Theta}(\psi,\Theta)\right\} = \mathbb{E}_{\Theta,\mathbf{Y}}\left\{\left(\Theta - \psi(\mathbf{Y})\right)^2\right\}. \tag{2.26}$$

En utilisant l'additivité de l'espérance, on peut exprimer l'erreur quadratique moyenne également comme somme de l'espérance du carré du biais et de la variance moyenne :

$$\mathbb{EQ}_{\Theta,\mathbf{Y}}(\psi) = \mathbb{E}_\Theta\left\{\mathbb{B}_{\mathbf{Y}|\Theta}(\psi,\Theta)^2 + \mathbb{V}_{\mathbf{Y}|\Theta}(\psi,\Theta)\right\} = \mathbb{E}_\Theta\left(\mathbb{B}_{\mathbf{Y}|\Theta}(\psi)^2\right) + \mathbb{V}_{\Theta,\mathbf{Y}}(\psi). \tag{2.27}$$

Remarque ⟨2.24⟩ *L'estimateur Espérance A Posteriori (Sect. 2.5.2) en tant qu'estimateur minimise l'erreur quadratique moyenne prise sous la loi jointe. Ainsi, les équations (2.26) et (2.27) définissent le risque bayésien attaché à la fonction de coût quadratique.*

10. Pour alléger les notations, on écrira pour le conditionnement $\mathbf{Y}\mid\Theta=\theta$ simplement $\mathbf{Y}\mid\theta$.

2.8.4 Biais, variance, erreur quadratique et l'estimateur parfait

Nous voulons qualifier l'estimateur $\psi \in \Psi$ en prenant en compte la qualité des estimations qu'il fournit. Pour cela, nous définissons l'estimateur parfait à travers des mentions que nous venons d'introduire.

Définition ⟨2.25⟩ *Pour un θ donné, l'estimateur $\psi : \Omega_{\mathbf{Y}} \to \Omega_\theta$, $\mathbf{Y} \mapsto \psi(\mathbf{Y})$ est appelé **parfait ponctuellement** en θ si l'erreur quadratique s'annule :*

$$\mathbb{E}\mathbb{Q}_{\mathbf{Y}|\Theta=\theta}(\psi,\theta) = \mathbb{E}_{\mathbf{Y}|\Theta=\theta}\left((\theta - \psi(\mathbf{Y}))^2\right) = 0.$$

Du fait que l'erreur quadratique est égale à la somme de la variance et du biais au carré, il en résulte que le biais et la variance s'annulent également :
- $\mathbb{B}_{\mathbf{Y}|\Theta=\theta}(\psi,\theta) = \mathbb{E}_{\mathbf{Y}|\Theta=\theta}\left(\theta - \psi(\mathbf{Y})\right) = 0$,
- $\mathbb{V}_{\mathbf{Y}|\Theta=\theta}(\psi,\theta) = \mathbb{E}_{\mathbf{Y}|\Theta=\theta}\left((\psi(\mathbf{Y}) - \mathbb{E}_{\mathbf{Y}|\Theta=\theta}(\psi(\mathbf{Y})))^2\right) = 0$.

Définition ⟨2.26⟩ *L'estimateur ψ est appelé **parfait** si pour tout θ, il est parfait ponctuellement.*

Corollaire ⟨2.27⟩ *L'estimateur parfait ψ possède une erreur quadratique moyenne nulle, $\mathbb{E}\mathbb{Q}_{\Theta,\mathbf{Y}}(\psi) = \mathbb{E}_{\Theta,\mathbf{Y}}\left((\Theta - \psi(\mathbf{Y}))^2\right) = 0$. Le biais moyen $\mathbb{B}_{\Theta,\mathbf{Y}}(\psi)$ et la variance moyenne $\mathbb{V}_{\Theta,\mathbf{Y}}(\psi)$ s'annulent également :*
- $\mathbb{B}_{\Theta,\mathbf{Y}}(\psi) = \mathbb{E}_{\Theta,\mathbf{Y}}\left(\Theta - \psi(\mathbf{Y})\right) = 0$,
- $\mathbb{V}_{\Theta,\mathbf{Y}}(\psi) = \mathbb{E}_{\Theta,\mathbf{Y}}\left((\psi(\mathbf{Y}) - \mathbb{E}_{\mathbf{Y}|\Theta}(\psi(\mathbf{Y})))^2\right) = 0$.

2.8.5 Coefficient de variation

Une mesure de performance couramment utilisée dans l'analyse de méthodes en protéomique, en probabilité et en statistiques est le *coefficient de variation*. Il s'agit d'une mesure normalisée de la dispersion de la distribution des estimations qui est définie comme rapport entre écart type et espérance de l'estimateur ψ pour un θ donné :

$$\mathbb{C}\mathbb{V}_{\mathbf{Y}|\theta}(\psi,\theta) = \frac{\left(\mathbb{V}_{\mathbf{Y}|\theta}(\psi,\theta)\right)^{1/2}}{\mathbb{E}_{\mathbf{Y}|\theta}\left[\psi(\mathbf{Y})\right]}. \tag{2.28}$$

Dans le cas d'une campagne expérimentale de type « rampe », *i.e.* l'acquisition de réplicats de tranches de valeur incrémentales (voir Sect. 3.6), cette définition nous permettra dans la suite de calculer un coefficient de variation par tranche de valeur vraie.

On voit que l'estimateur parfait ponctuellement au point θ a un coefficient de variation nul puisque sa variance $\mathbb{V}_{\mathbf{Y}|\theta}(\psi,\theta)$ est nulle. La réciproque cependant est fausse ! Il suffit de considérer l'estimateur $\psi(\mathbf{Y}) = c$ pour tout $\mathbf{Y} \in \Omega_{\mathbf{Y}}$, c étant une valeur arbitraire constante. Le coefficient de variation est bien sûr nul car la variance de l'estimateur l'est, mais l'estimateur n'est pas parfait. En effet, cette mesure renseigne uniquement sur la *dispersion des estimations*, mais pas sur l'exactitude des estimations par rapport aux valeurs vraies.

Il est important de souligner ici qu'il s'agit d'une mesure de performance en fonction de la variance de l'*estimation* et non en fonction de la vérité. Ainsi, deux estimateurs ψ_1 et ψ_2 associés à des vérités distinctes θ_1 et θ_2 respectivement ont le même coefficient de variation si les estimations sont les mêmes (à des permutations près). Alors que les vérités ne sont pas égales, *i.e.* un estimateur est plus biaisé que l'autre, ils affichent la même dispersion.

Le coefficient de variation moyen se calcule en prenant l'espérance sous la loi de θ du coefficient de variation :

$$\mathbb{C}\mathbb{V}_{\Theta,\mathbf{Y}}(\psi) = \mathbb{E}_\Theta\left[\mathbb{C}\mathbb{V}_{\mathbf{Y}|\theta}(\psi,\theta)\right] = \mathbb{E}_\Theta\left[\frac{\left(\mathbb{V}_{\mathbf{Y}|\theta}(\psi,\theta)\right)^{1/2}}{\mathbb{E}_{\mathbf{Y}|\theta}\left[\psi(\mathbf{Y})\right]}\right]. \tag{2.29}$$

Notons que le coefficient de variation moyen s'annule pour l'estimateur parfait puisqu'il est parfait pour toute valeur de θ, ce qui implique que la variance s'annule en tout point θ, mais la réciproque est, comme pour le CV ponctuel, fausse.

Finalement, le coefficient de variation global s'obtient comme l'espérance sous la loi de θ dans la variance du numérateur et dans l'espérance du dénominateur :

$$\mathbb{CV}^{\text{global}}_{\Theta,\mathbf{Y}}(\psi) = \frac{\left(\mathbb{V}_{\Theta,\mathbf{Y}}(\psi)\right)^{1/2}}{\mathbb{E}_{\Theta,\mathbf{Y}}\left[\psi(\mathbf{Y})\right]}. \tag{2.30}$$

2.8.6 Droite de régression

La **droite de régression simple** propose d'utiliser un modèle affine pour la prédiction de valeurs. Elle est souvent utilisée pour l'évaluation d'une méthode de quantification, par exemple en protéomique quantitative pour répondre à la question de la fiabilité d'un quantifieur.

La droite de régression lie d'une manière affine les quantités vraies θ aux quantités estimées $\psi(\mathbf{Y})$ par une droite et s'exprime par la relation $f(\theta) = \alpha + \beta\theta$ où l'ordonnée à l'origine est α et la pente β.

En fonction du critère d'erreur choisi, la droite peut avoir de coefficients de régressions différents ; nous allons nous intéresser aux critères des moindres carrés. Les sous-sections suivantes n'énoncent que des résultats ; le lecteur peut cependant suivre les calculs détaillés dans l'annexe Ann. A.3.

Coefficients de la droite de régression

Nous nous donnons un estimateur quelconque ψ et cherchons à minimiser l'espérance de l'écart quadratique entre l'estimation $\psi(\mathbf{Y})$ et la fonction affine $f(\theta) = \alpha + \beta\theta$ sous la distribution jointe, par rapport aux paramètres de régression α et β. Cette espérance s'écrit

$$\mathbb{J}_{\theta,\mathbf{Y},\psi}(\alpha,\beta) = \mathbb{E}_{\Theta,\mathbf{Y}}((\psi(\mathbf{Y}) - f(\Theta))^2) = \mathbb{E}_{\Theta,\mathbf{Y}}((\psi(\mathbf{Y}) - \alpha - \beta\Theta)^2), \tag{2.31}$$

et atteint son minimum par rapport à l'estimateur ψ en $(\bar{\alpha},\bar{\beta})$ tels que

$$\bar{\alpha} = \mathbb{E}_{\mathbf{Y}}(\psi(\mathbf{Y})) - \bar{\beta}\mathbb{E}_{\Theta}(\Theta). \tag{2.32}$$

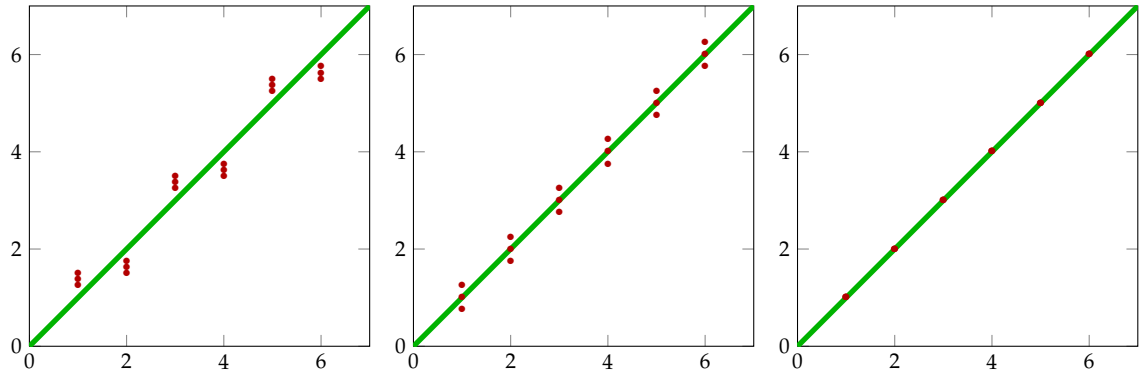
$$\bar{\beta} = \frac{\mathbb{E}_{\Theta,\mathbf{Y}}(\Theta\cdot\psi(\mathbf{Y})) - \mathbb{E}_{\Theta}(\Theta)\mathbb{E}_{\mathbf{Y}}(\psi(\mathbf{Y}))}{\mathbb{E}_{\Theta}(\Theta^2) - \mathbb{E}_{\Theta}(\Theta)^2} = \frac{\mathbb{C}_{\Theta,\mathbf{Y}}(\Theta,\psi(\mathbf{Y}))}{\mathbb{V}_{\Theta}(\Theta)}. \tag{2.33}$$

La fonction affine optimale, appelée *prédiction*, est donnée par $\bar{f}(\theta) = \bar{\alpha} + \bar{\beta}\theta$.

Tandis que le paramètre $\bar{\beta}$ quantifie le rapport entre la covariance des variables Θ et $\mathbf{Y}$ et la variance de la variable explicative Θ, le paramètre $\bar{\alpha}$ donne la différence entre la moyenne des estimations et la moyenne des entrées pondérée par $\bar{\beta}$.

Comment ces paramètres optimaux nous renseignent-ils sur la qualité de l'estimateur ? Pour cela, considérons le cas où $\bar{\alpha} = 0$ **et** $\bar{\beta} = 1$ conjointement, *i.e.* $\bar{f}(\theta) = \theta$. Dans ce cas précis, l'ordonnée à l'origine est nulle et nous n'avons pas de décalage systématique des valeurs estimées (en moyenne), l'espérance des estimations égalant $\bar{\beta} = 1$ fois l'espérance des entrées : le biais moyen est nul. Ensuite, la covariance entre Θ et $\mathbf{Y}$ compense la variance de Θ. Cependant, uniquement avec ces deux paramètres, nous ne pouvons qualifier l'estimateur parfait car cette fonction optimale est attachée à plusieurs nuages de points (voir Fig. 2.3). C'est la raison pour laquelle nous devons prendre en considération une autre caractéristique de la régression simple : le coefficient de détermination.

Fig. 2.3 Schématisation de trois nuages de points conduisant à une régression par l'identité



(a) Nuage alternant, biais non nuls, biais moyen nul, variance non nulle, $R^2_{\alpha,\beta}$ petit

(b) Nuage centré, biais nuls, biais moyen nul, variances non nulles, $R^2_{\alpha,\beta}$ élevé

(c) Nuage réduit à des points, biais nuls, biais moyen nul, variance nulle, $R^2_{\alpha,\beta} = 1$

Coefficient de détermination

On introduit ici le *coefficient de détermination* $R^2_{\alpha,\beta} \in [0,1]$ associé à la droite $f(\theta) = \alpha + \beta\theta$: il permet d'évaluer le degré d'association entre les valeurs prédites $f(\theta)$ par la droite et les estimations $\psi(\mathbf{Y})$.

Définition ⟨2.28⟩ *Le coefficient de détermination $R^2_{\alpha,\beta}$ est défini de la manière suivante :*

$$R^2_{\alpha,\beta} = 1 - \frac{\mathbb{E}_{\Theta,\mathbf{Y}}\left((\psi(\mathbf{Y}) - f(\Theta))^2\right)}{\mathbb{E}_{\mathbf{Y}}\left((\psi(\mathbf{Y}) - \mathbb{E}_{\mathbf{Y}}(\psi(\mathbf{Y})))^2\right)} = 1 - \frac{\mathbb{J}_{\Theta,\mathbf{Y},\psi}(\alpha,\beta)}{\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))}. \tag{2.34}$$

Le $R^2_{\alpha,\beta}$ juge de la qualité de la représentation des points par la droite de régression $f(\theta) = \alpha + \beta\theta$. Si elle passe exactement par tous les points (voir Fig. 2.3(c)), alors $R^2_{\alpha,\beta} = 1$ ce qui indique que toute la variation des estimations est expliquée par les valeurs vraies. On peut dire que plus le $R^2_{\alpha,\beta}$ est élevé, plus les valeurs vraies contiennent d'information sur les estimations, les cas limites étant évidemment $R^2_{\alpha,\beta} = 0$ (θ ne contient aucune information sur $\psi(\mathbf{Y})$) et $R^2_{\alpha,\beta} = 1$ (θ contient toute l'information sur $\psi(\mathbf{Y})$ comme indiqué précédemment). Dans le contexte de la prédiction, plus $R^2_{\alpha,\beta}$ s'approche de 1, plus le modèle explique bien les estimations (ce qui ne veut pas dire que le modèle est juste) ; plus $R^2_{\alpha,\beta}$ est proche de 0, moins l'apport du modèle est grand.

Un calcul simple (voir l'Ann. A.4) montre ensuite la relation entre le coefficient de détermination $R^2_{\bar{\alpha},\bar{\beta}}$ et l'écart quadratique moyen $\mathbb{J}_{\theta,\mathbf{Y},\psi}(\bar{\alpha},\bar{\beta})$ respectivement pour la droite de régression optimale. L'écart quadratique moyen admet une borne inférieure qui est atteinte pour les coefficients de régression optimale, *i.e.* $\mathbb{J}_{\theta,\mathbf{Y},\psi}(\bar{\alpha},\bar{\beta}) = \mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))(1 - R^2_{\bar{\alpha},\bar{\beta}})$. D'avoir pu borné le critère d'écart quadratique par rapport à la droite de régression veut notamment dire que le coefficient de détermination admet une borne supérieure $R^2_{\bar{\alpha},\bar{\beta}} \leq 1$ atteinte par la régression optimale.

Par cette identification dont les détails sont rapportés en annexe (Ann. A.4), nous pouvons écrire le coefficient de détermination de la droite de régression optimale dans le cas d'une régression linéaire simple comme rapport du carré de la covariance entre les

variables et du produit des variances respectives des variables :

$$R^2_{\bar{\alpha},\bar{\beta}} = \frac{\mathbb{C}_{\Theta,\mathbf{Y}}(\Theta, \psi(\mathbf{Y}))^2}{\mathbb{V}_{\Theta}(\Theta) \cdot \mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))}. \tag{2.35}$$

Cette expression met en évidence que le coefficient de détermination de la droite de régression optimale $R^2_{\bar{\alpha},\bar{\beta}}$ est égal au carré de la valeur de la corrélation du nuage de points, aussi appelée *corrélation de Pearson*[v] [104].

La covariance étant symétrique, l'expression du $R^2_{\bar{\alpha},\bar{\beta}}$ dans Éqn. (2.35) peut être interprétée comme étant le produit des pentes de deux droites de régression complémentaires :

1. le premier facteur est $\bar{\beta}_{\theta \div \mathbf{Y}} = \mathbb{C}_{\Theta,\mathbf{Y}}(\Theta, \psi(\mathbf{Y}))/\mathbb{V}_{\Theta}(\Theta)$: θ est la variable explicative, $\psi(\mathbf{Y})$ est la variable régressée,
2. le deuxième facteur est $\bar{\beta}_{\mathbf{Y} \div \theta} = \mathbb{C}_{\Theta,\mathbf{Y}}(\Theta, \psi(\mathbf{Y}))/\mathbb{V}_{\mathbf{Y}}(\mathbf{Y})$: $\psi(\mathbf{Y})$ est la variable explicative, θ est la variable régressée.

La régression est donc parfaite si la pente du premier est l'inverse de la pente du deuxième.

L'estimateur parfait et la régression simple

Plus haut, nous avons mis en relation les coefficients de la fonction affine optimale $\bar{f}(\theta) = \bar{\alpha} + \bar{\beta}\theta$ et les mesures d'erreur associées à un estimateur ψ. Nous y avons vu que la fonction identité seule ne suffisait pas pour décrire l'estimateur parfait. Pour compléter l'interprétation, nous avons proposé la définition du coefficient de détermination $R^2_{\alpha,\beta}$. Un coefficient de détermination de valeur 1 indique que l'écart des estimations par rapport à la droite de régression est nulle, mais pas par rapport à l'entrée. En effet, n'importe quelle droite de régression $f(\theta) = \alpha + \beta\theta$ peut avoir un $R^2_{\alpha,\beta}$ égalant 1 sans être l'estimateur parfait.

C'est la combinaison des deux notions « identité » et « $R^2_{\alpha,\beta} = 1$ » qui permet d'énoncer le lien entre *mesures d'erreur d'un estimateur parfait* (voir Sect. 2.8.4) et *coefficient de la droite de régression*.

Proposition ⟨2.29⟩ *L'estimateur ψ est parfait si et seulement si la droite de régression $f(\theta) = \alpha + \beta\theta$ associé au nuage de points des estimations est l'identité $f(\theta) = \theta$ et que le coefficient de détermination vaut $R^2_{\alpha,\beta} = 1$, i.e. les coefficients de régression valent $(\alpha, \beta, R^2_{\alpha,\beta}) = (0, 1, 1)$.*

Preuve
Nous donnons l'idée de la preuve, les détails se trouvent en annexe, voir Ann. A.5.

Le sens « parfait $\Rightarrow$ $(\alpha, \beta, R^2_{\alpha,\beta}) = (0, 1, 1)$ » est immédiat en utilisant la Déf. ⟨2.26⟩ et le fait que la somme de termes non négatifs ne peut s'annuler que si les termes eux-mêmes sont nuls.

Le sens « parfait $\Leftarrow$ $(\alpha, \beta, R^2_{\alpha,\beta}) = (0, 1, 1)$ » s'obtient en cherchant à identifier les moments des lois pour Θ, pour $\mathbf{Y}$ et pour le couple $(\Theta, \mathbf{Y})$. On se sert pour cela des Éqs. (2.32), (2.33) et (2.35).■

2.8.7 Erreur de classification

On s'intéresse aux performances de classification. Soit $\mathcal{C} = \Omega_c$ l'ensemble des classes, et soit $M = \mathrm{card}(\mathcal{C})$ sa cardinalité. Ensuite, soit $\psi : \Omega_{\mathbf{Y}} \to \mathcal{C}$ le classifieur qui associe à une donnée $\mathbf{Y}$ une classe $\psi(\mathbf{Y})$.

Soit $c \in \mathcal{C}$ une classe fixée. On définit alors l'erreur de classification par la fonction de coût $L : \mathcal{C} \times \mathcal{C} \to \mathbb{R}_+$:

$$L(c, \psi(\mathbf{Y})) = \begin{cases} 0 & \text{si } \psi(\mathbf{Y}) = c, \\ \ell_{c,\tilde{c}} & \text{si } \psi(\mathbf{Y}) = \tilde{c} \neq c. \end{cases} \tag{2.36}$$

Le paramètre $\ell_{c,\tilde{c}}$ quantifie l'impact de la mauvaise classification d'un échantillon de classe vraie c dans la classe $\tilde{c}$.

L'erreur de classification est alors l'espérance de la fonction de coût sous la loi de $\mathbf{Y} \mid C = c$:

$$\mathrm{R}_{\mathbf{Y}|c}(c, \psi) = \mathbb{E}_{\mathbf{Y}|c}\left(\mathrm{L}(c, \psi(\mathbf{Y}))\right). \tag{2.37}$$

Cette définition correspond à celle du risque d'un classifieur. L'erreur moyenne de classification se fait tout naturellement en prenant l'espérance du risque :

$$\begin{aligned} \mathrm{R}_{C,\mathbf{Y}}(\psi) &= \mathbb{E}_C\left(\mathrm{R}_{\mathbf{Y}|C}(C, \psi(\mathbf{Y}))\right) \\ &= \mathbb{E}_{C,\mathbf{Y}}(\mathrm{L}(C, \psi(\mathbf{Y}))) ; \end{aligned} \tag{2.38}$$

on parlera aussi de risque moyen.

Exemple ⟨2.30⟩ *Prenons le cas* $\mathsf{M} = 2$ *avec* $\mathcal{C} = \{\mathrm{S}, \mathrm{M}\}$ *— hypothèse nulle : l'individu est « sain » et hypothèse alternative : l'individu est « malade » — et un coût* 0–1, *i.e. la mauvaise détection d'une classe est quantifiée par* 1, *quelle que soit la classe vraie. Nous distinguons quatre cas :*

Vrai positif : *l'individu malade a été classé malade.*

Vrai négatif : *l'individu sain a été classé sain.*

Faux positif : *l'individu sain a été classé malade.*

Faux négatif : *l'individu malade a été classé sain.*

Ainsi, le taux de faux positifs (erreur type I) pour le classifieur ψ *est équivalent au risque attaché à la classe* $C = \mathrm{S}$, $\mathrm{R}_{\mathbf{Y}|C=\mathrm{S}}(C = \mathrm{S}, \psi(\mathbf{Y}) = \mathrm{M})$, *qui quantifie la probabilité de classer un individu sain dans la classe des malades.*

Le taux de faux négatifs (erreur type II) équivaut au risque attaché à la classe $C = \mathrm{M}$, $\mathrm{R}_{\mathbf{Y}|C=\mathrm{M}}(C = \mathrm{M}, \psi(\mathbf{Y}) = \mathrm{S})$ *quantifiant les mauvais classements d'individus malades dans l'ensemble d'individus sains.*

Finalement, le taux de mauvaise classification est exprimé par le risque moyen $\mathrm{R}_{C,\mathbf{Y}}(\psi)$. *Nous pouvons résumer ceci dans un* tableau des risques *bi-classe comme suit :*

vérité / ***estimation***	individu sain	individu malade
classé sain	$1 - \mathrm{R}_{\mathbf{Y}\|C=\mathrm{S}}(C = \mathrm{S}, \psi(\mathbf{Y}) = \mathrm{M})$ *vrai négatif*	$\mathrm{R}_{\mathbf{Y}\|C=\mathrm{M}}(C = \mathrm{M}, \psi(\mathbf{Y}) = \mathrm{S})$ *faux négatif*
classé malade	$\mathrm{R}_{\mathbf{Y}\|C=\mathrm{S}}(C = \mathrm{S}, \psi(\mathbf{Y}) = \mathrm{M})$ *faux positif*	$1 - \mathrm{R}_{\mathbf{Y}\|C=\mathrm{M}}(C = \mathrm{M}, \psi(\mathbf{Y}) = \mathrm{S})$ *vrai positif*

2.8.8 Divergence de Kullback-Leibler

La *divergence de Kullback-Leibler*[vi][vii] permet de mesurer la « distance » entre deux distributions. Mathématiquement, il ne s'agit pas d'une fonction de distance car ni la propriété de symétrie ↓11 ni l'inégalité triangulaire ne sont vérifiées.

11. Pour pallier le manque de symétrie, on trouve souvent aussi la divergence de Kullback-Leiber symétrisée $\mathrm{KL}_{\mathrm{sym}}(p, q) = \mathrm{KL}\,(p\|q) + \mathrm{KL}\,(q\|p)$.

Définition ⟨2.31⟩ *Soient p et q deux densités de distributions du paramètre θ. La divergence de Kullback-Leibler est définie par l'expression suivante [21, Sect. 1.6.1] :*

$$\mathrm{KL}\,(p\|q) = \int_{\Omega_\theta} p(\theta) \log \frac{p(\theta)}{q(\theta)}\,\mathrm{d}\theta. \tag{2.39}$$

Mentionnons certaines propriétés de cette mesure.

(P1) **Non négativité**. La divergence de Kullback-Leibler est positive ou nulle quelles que soient les distributions p et q : $\mathrm{KL}\,(p\|q) \geq 0$.

(P2) **Unicité**. La divergence de Kullback-Leibler de deux distributions p et q s'annule si et seulement si $p = q$.

(P3) **Additivité**. Soient le paramètre $\theta = (\theta_1, \theta_2)$ et les distributions p, q séparables, *i.e.* $p(\theta_1, \theta_2) = p_1(\theta_1)p_2(\theta_2)$ et $q(\theta_1, \theta_2) = q_1(\theta_1)q_2(\theta_2)$. Alors, la divergence jointe est la somme des divergences séparées, $\mathrm{KL}\,(p\|q) = \mathrm{KL}\,(p_1\|q_1) + \mathrm{KL}\,(p_2\|q_2)$.

(P4) **Invariance aux transformations**. Pour la transformation $\phi = h(\theta)$, la divergence de Kullback-Leibler entre deux distributions p et q est invariante :

$$\begin{aligned}\mathrm{KL}\,(p\|q) &= \int p(\theta) \log \frac{p(\theta)}{q(\theta)}\,\mathrm{d}\theta \\ &= \int p(h(\theta))h'(\theta) \log \frac{p(h(\theta))h'(\theta)}{q(h(\theta))h'(\theta)}\,\mathrm{d}\theta = \int p(\phi) \log \frac{p(\phi)}{q(\phi)}\,\mathrm{d}\phi.\end{aligned}$$

Une transformation de paramètre (comme par exemple le passage au logarithme) induit une transformation de lois associée, mais elle n'affecte pas la « distance » entre ces deux lois, elle est constante. La divergence de Kullback-Leibler peut être considérée comme une quantité géométrique qui est indépendante du repère choisi [42, Sect. 4.2].

(P5) **Distributions normales**. La divergence de Kullback-Leibler pour deux distributions normales de dimension N, $p(\theta) = \mathcal{N}(\theta; \boldsymbol{m}_p, \boldsymbol{\Gamma}_p)$ et $q(\theta) = \mathcal{N}(\theta; \boldsymbol{m}_q, \boldsymbol{\Gamma}_q)$, a la forme [42, Sect. 4.2]

$$\mathrm{KL}\,(p\|q) = \frac{1}{2}\Big(\underbrace{\mathrm{tr}\left(\boldsymbol{\Gamma}_q^{-1}\boldsymbol{\Gamma}_p\right)}_{\text{(i)}} + \underbrace{\left\|\boldsymbol{m}_q - \boldsymbol{m}_p\right\|_{\boldsymbol{\Gamma}_q}^2}_{\text{(ii)}} - \underbrace{\log\left(\left|\boldsymbol{\Gamma}_q^{-1}\boldsymbol{\Gamma}_p\right|\right)}_{\text{(iii)}} - \underbrace{N}_{\text{(iv)}}\Big).$$

Cette expression fait intervenir (dans l'ordre)

(i) la somme des valeurs propres du rapport des précisions,

(ii) la norme euclidienne de la différence des moyennes déformée par la matrice de précision de la distribution q [12],

(iii) le logarithme du produit des valeurs propres du rapport des précisions (l'opérateur $|\cdot|$ désigne le déterminant d'une matrice), ou autrement dit, en profitant des propriétés du logarithme matriciel, la somme des valeurs propres du logarithme du quotient matriciel des précisions, et

(iv) la dimension du paramètre θ.

12. Pour $\boldsymbol{\Gamma}_p = \boldsymbol{\Gamma}_q$, cette norme devient la norme de Mahalanobis par rapport à la matrice de covariance $\boldsymbol{\Gamma}_q^{-1}$ [43, Sect. 6.2.5], [42, Sect. 4.2].

2.9 Outils fréquents pour le calcul bayésien

Le calcul des lois *a posteriori* peut être fastidieux, voire même infaisable analytiquement quand les lois et les vraisemblances deviennent compliquées et que les données ou les inconnues sont de grande taille. Dans ce cas, nous avons recours à des méthodes numériques pour calculer une approximation. Deux méthodes majeures de référence sont les méthodes « Monte Carlo Chaîne de Markov » [viii][ix][x][xi] (*Monte Carlo Markov Chain*, MCMC) et « Bayésien variationnel » (*Variational Bayes*, VB) que nous introduirons dans cette section.

2.9.1 Markov-Chain Monte-Carlo

L'ouvrage complet [21] livre beaucoup de détails sur le fonctionnement des méthodes MCMC et des échantillonneurs. Le contenu des sections suivantes y est excellemment bien expliqué.

Les méthodes MCMC ([2, Sect. 6.3], [3, Sect. 11.2]) sont un couplage des méthodes *Monte Carlo*, *i.e.* résolution d'un problème d'intégration par échantillonnage stochastique, et les *Chaînes de Markov* d'ordre 1, *i.e.* des chaînes qui déterminent aléatoirement une valeur en fonction de la précédente uniquement. La stratégie MCMC consiste à définir une chaîne de Markov irréductible et apériodique qui a la distribution cible, donc la loi *a posteriori* totale, comme distribution stationnaire. Une fois dans un état stationnaire, les échantillons tirés permettent

- d'approcher la loi (par considération de l'histogramme),
- de calculer les estimateurs (par moyennage des échantillons pour l'estimateur EAP, prise de l'échantillon le plus souvent échantillonné pour MAP, ...),
- de calculer les moments de la loi et ainsi entre autre la variance (par moyennage des échantillons élevés à la puissance souhaitée),
- de marginaliser des paramètres (par projection sur les paramètres d'intérêt uniquement), *et cetera*.

Afin d'avoir des échantillons, nous avons besoin d'un échantillonneur qui en propose à chaque itération ce qui est le sujet des sous-sections suivantes.

Remarque ⟨2.32⟩ (Temps de chauffe) *Les premiers échantillons d'un algorithme*[xii] *d'échantillonnage stochastique de type MCMC sont fortement influencés par l'initialisation. Il faut donc attendre K_0 itérations pour que les échantillons soient distribués sous la loi cible. Ce temps d'attente est appelé* ***temps de chauffe*** *et peut varier d'une application à l'autre. [3, Sect. 11.6] propose par exemple d'omettre la première moitié des échantillons et de les considérer comme échantillons de chauffe. On peut aussi démarrer plusieurs chaînes, initialisées à des valeurs différentes, pour surveiller la convergence. Cette dernière est atteinte quand toutes les chaînes échantillonnent la même loi cible. Ceci peut être décidé par l'étude des variances intra- et inter-chaîne, comme propose [3, Sect. 11.6].*

L'échantillonnage peut être long, la convergence lente, les calculs lourds malgré tous les efforts. Pour ne pas devoir lancer plusieurs chaînes à chaque fois que l'on s'apprête à faire une inversion en utilisant les méthodes MCMC, il suffit d'étudier un cas typique et d'en extraire le temps de chauffe. Ce résultat peut être utilisé dans les prochaines utilisations d'un MCMC pour trouver l'itération à laquelle le temps de chauffe se termine.

Échantillonnage sous une loi et générateur d'échantillons

Avant de décrire les outils, nous expliquons ce que nous appelons échantillonnage ou échantillonnage stochastique, termes récurrents dans tout le document.

L'échantillonnage d'une variable aléatoire sous sa distribution est la sélection aléatoire d'une valeur parmi les valeurs d'un ensemble spécifique. Cet ensemble (fini ou infini) est déterminé par la distribution de la variable aléatoire.

Plus on tire d'échantillons sous la distribution donnée, plus le sous-ensemble des échantillons est semblable à l'ensemble. Autrement dit, plus on a d'échantillons de la loi, plus son histogramme normalisé ressemble à la loi de probabilité échantillonnée. Si on est dans un cas de loi de probabilité continue, l'échantillonnage est alors un moyen d'approcher la densité par un ensemble discret, avec une meilleure approximation quand le nombre d'échantillons augmente.

Quand la loi de probabilité a une forme usuelle comme par exemple une loi uniforme, normale, gamma, bêta, laplacienne, poissonnienne, bi- ou multinomiale, il existe dans la plupart des logiciels de calcul numérique et statistique (MATLAB, Octave, FreeMat, SciLab, R, S-plus, Sage, Maple, ...) des générateurs d'échantillons. En guise d'exemple, MATLAB propose une boîte à outil payante, *Statistics Toolbox*, qui permet entre autre d'échantillonner sous des lois usuelles [105] ; il existe la boîte à outil *Lightspeed* [106] gratuite qui permet de faire la même chose et qui a été conçue pour MATLAB, mais qui est compatible avec Octave. Le logiciel statistique libre R a également intégré un grand nombre de générateurs d'échantillonneurs.

Dans certains cas, le seul générateur existant est celui d'une loi uniforme entre 0 et 1, par exemple dans des environnements de programmation C, C++, Java et pour la programmation des générateurs ci-dessus. Par des méthodes de transformation, d'inversion ou de rejet selon la distribution, on peut transformer un échantillon d'une loi uniforme entre 0 et 1 en un échantillon sous la loi souhaitée. La mise en œuvre informatique est discutée dans [44] mais semble dépasser le cadre de ce document.

Metropolis-Hastings

Une construction populaire d'un MCMC remonte aux années 1950 [45] où le physicien gréco-américain Nicholas Metropolis[xiii][viii] présente un algorithme échantillonnant sous une loi cible à l'aide d'une loi de proposition symétrique. Dans les années 1970 [46], le mathématicien canadien W. Keith Hastings[xiv] étend ces travaux et développe l'algorithme de Metropolis-Hastings (MH) [3, Sect. 11.4]. Son principe est simple. Nous souhaitons échantillonner sous une loi cible $\varphi(\theta)$, par exemple la loi *a posteriori* $\mathrm{p}(\theta \mid \mathbf{Y})$. À partir d'une valeur de départ $\theta^{(0)}$ qui est dans le support de la loi, nous démarrons une méthode itérative. On note (k) l'indice de l'itération courante, le dernier échantillon du paramètre est donc $\theta^{(k-1)}$. Nous allons *proposer* une nouvelle valeur θ^{p}, issue d'une *loi de proposition* $\pi(\theta \mid \theta^{(k-1)})$ qui dépend du dernier échantillon $\theta^{(k-1)}$ (caractère Chaîne de Markov).

La valeur proposée est ensuite acceptée comme valeur de $\theta^{(k)}$ avec une probabilité 1 si elle est plus probable pour la loi cible, et avec une probabilité $\delta < 1$ sinon. L'expression de δ est donnée par le rapport des densités cibles pondéré par les densités de proposition :

$$\delta = \min\left(1, \frac{\varphi(\theta^{\mathrm{p}})}{\varphi(\theta^{(k-1)})} \bigg/ \frac{\pi(\theta^{\mathrm{p}} \mid \theta^{(k-1)})}{\pi(\theta^{(k-1)} \mid \theta^{\mathrm{p}})}\right) \tag{2.40}$$

Remarque ⟨2.33⟩ *Quand la loi cible est la loi a posteriori, on a*

$$\begin{aligned}\delta &= \min\left(1, \frac{\mathrm{p}(\theta^{\mathrm{p}} \mid \mathbf{Y})}{\mathrm{p}(\theta^{(k-1)} \mid \mathbf{Y})} \bigg/ \frac{\pi(\theta^{\mathrm{p}} \mid \theta^{(k-1)})}{\pi(\theta^{(k-1)} \mid \theta^{\mathrm{p}})}\right) \\ &= \min\left(1, \frac{\mathrm{p}(\theta^{\mathrm{p}})\,\mathrm{p}(\mathbf{Y} \mid \theta^{\mathrm{p}})}{\mathrm{p}(\theta^{(k-1)})\,\mathrm{p}(\mathbf{Y} \mid \theta^{(k-1)})} \bigg/ \frac{\pi(\theta^{\mathrm{p}} \mid \theta^{(k-1)})}{\pi(\theta^{(k-1)} \mid \theta^{\mathrm{p}})}\right).\end{aligned} \tag{2.41}$$

Nous pouvons distinguer des cas spéciaux.

1. *Si la probabilité* a priori *est égale pour toutes les valeurs (loi uniforme), la probabilité d'acceptation est égale au quotient du rapport de vraisemblances et du rapport de densités de proposition π.*
2. *Si la loi de proposition est symétrique, i.e. $\pi(\theta^{\mathrm{p}}, \theta^{(k-1)}) = \pi(\theta^{(k-1)}, \theta^{\mathrm{p}})$, alors la probabilité d'acceptation est le rapport des densités cibles, donc le rapport des densités* a posteriori.
3. *Si la loi* a priori *est uniforme et la densité de proposition symétrique, alors la probabilité d'acceptation n'est rien d'autre que le rapport des vraisemblances.*
4. *Si, enfin, la densité de proposition est égale à la distribution* a priori, *alors la probabilité d'acceptation est également le rapport des vraisemblances.*

Remarque ⟨2.34⟩ *Si π est symétrique, i.e. $\pi(\theta \mid \theta') = \pi(\theta' \mid \theta)$, alors on est dans le cas de l'algorithme de Metropolis.*

L'article [47] étudie les lois de proposition et leurs comportements, comme par exemple la vitesse de convergence et le temps de calcul. En effet, suivant la loi de proposition, le calcul par itération peut être peu coûteux, mais la convergence très lente ; ou bien le calcul par itération peut être très cher en temps (p.ex. calcul de matrices hessiennes), mais la convergence est très rapide car les propositions sont orientées vers la zone de forte probabilité. Nous présenterons ici les deux lois de proposition les plus simples [3, Sect. 11.9].

Metropolis-Hastings Indépendant

Metropolis-Hastings Indépendant (MHI) est caractérisé par sa loi de proposition qui ne dépend pas de la valeur courante : $\pi(\theta^{\mathrm{p}} \mid \theta^{(k-1)}) = \pi(\theta^{\mathrm{p}})$. Un choix possible et souvent fait dans ce cas est une loi uniforme sur l'espace du paramètre. Ainsi, le rapport des densités de proposition vaut 1 et nous nous retrouvons avec un algorithme de Metropolis. Le calcul est très rapide, et tout l'espace du paramètre est exploré. Cependant, les propositions sont souvent éloignées de la zone de forte probabilité, ce qui fait que l'algorithme a besoin de beaucoup d'itérations pour converger.

Un autre choix courant est celui de la loi *a priori* du paramètre θ pour la loi de proposition. La dernière remarque concernant la distance par rapport à la zone de forte probabilité tient toujours pour des lois larges.

Metropolis-Hastings à Marche Aléatoire

L'algorithme de Metropolis-Hastings à Marche Aléatoire (MHMA) part du principe que, si une valeur a été choisie précédemment, c'est qu'il y a des chances qu'une autre valeur à forte probabilité soit dans un voisinage proche. La densité de proposition exprime ainsi une marche aléatoire, partant du point $\theta^{(k-1)}$. On peut par exemple choisir une densité uniforme sur un petit intervalle avec moyenne $\theta^{(k-1)}$, ou une densité de proposition normale de moyenne $\theta^{(k-1)}$ et de précision γ_{MA}. Notons que cette dernière densité est symétrique, $\mathcal{N}(\theta^{\mathrm{p}}; \theta^{(k-1)}, \gamma_{\mathrm{MA}}) = \mathcal{N}(\theta^{(k-1)}; \theta^{\mathrm{p}}, \gamma_{\mathrm{MA}})$, ce qui fait que la probabilité d'acceptation a la forme simple de l'algorithme de Metropolis.

Alors que MHI ne nécessite en général pas de paramétrage explicite, MHMA a le paramètre γ_{MA}. C'est lui qui définit le voisinage de la nouvelle proposition par rapport à la valeur courante. L'utilisateur doit choisir la valeur pour γ_{MA} de telle manière que

- les propositions s'éloignent suffisamment de la valeur courante pour permettre une bonne exploration de l'espace des paramètres et pour améliorer la probabilité *a posteriori*,

– les propositions ne s'éloignent pas trop de la valeur courante pour ne pas être rejetée incessamment car issues d'une région de l'espace improbable.

[3, Sect. 11.10] propose de paramétrer l'algorithme de telle façon que la proportion de propositions acceptées soit 40% environ (selon la dimension du vecteur de paramètres).

Remarque ⟨2.35⟩ *Soit θ un vecteur de I paramètres. Dans ce cas et quelle que soit la méthode* MH *qu'on utilise, il y a un autre choix à faire qui impacte la vitesse de convergence de l'échantillonnage. On peut soit échantillonner le vecteur θ sur toutes les composantes à la fois, c'est-à-dire on tire en un seul échantillonnage un vecteur de valeurs pour toutes les composantes. On l'appellera* ***simulation parallèle****. Soit, on échantillonne le vecteur θ sur les composantes les unes après les autres. Ceci nécessitera l'échantillonnage sous la loi de θ_i pour la composante i sachant les autres composantes $\theta_{\{1,\ldots,I\}\setminus\{i\}}$. On y référera sous le nom de* ***simulation séquentielle****. Remarquons que la simulation séquentielle n'effectue rien d'autre qu'une structure de Gibbs au niveau du paramètre, voir ci-dessous.*

Il y a des pour et des contre. Généralement, la simulation séquentielle converge un peu plus vite. Dans la simulation parallèle, toutes les composantes évoluent en même temps ou n'évoluent pas du tout, contrairement à la simulation séquentielle où, en ne considérant qu'une composante, le vecteur évolue à chaque acceptation d'une composante.

Inversement, la simulation parallèle a souvent besoin de moins de temps de calcul par itération car il suffit d'évaluer globalement une *probabilité d'acceptation δ_{global} et non un δ_i par composante, soit un nombre total de I probabilités d'acceptation par itération.*

Le choix doit être fait au cas par cas, considérant la convergence souhaitée, le temps de calcul à investir, l'espace (voire les espaces) des paramètres, l'informativité des lois a priori et cetera.

Structure et échantillonneur de Gibbs

Un des concepts les plus attrayants pour l'échantillonnage est la **structure de Gibbs**[xv] ([3, Sects. 11.3 et 11.8]). Elle permet de passer du problème d'échantillonnage d'une loi compliquée (*e.g.* la loi *a posteriori* totale) à une succession de sous-problèmes d'échantillonnage de lois plus simples. Il s'agit là des **lois *a posteriori* conditionnelles** pour chacun des paramètres ou pour des groupes de paramètres pour lesquels il est « facile » d'échantillonner. (Notons que le terme « facile » est subjectif et extensible ...)

La simplification de l'échantillonnage global à des échantillonnages locaux est encore une fois renforcée par la structure hiérarchique s'il y en a une. Souvenons-nous que la justification d'un modèle hiérarchique bayésien (Sect. 2.7) est l'indépendance conditionnelle des paramètres (Sect. 2.7.2). Si le vecteur de paramètres $\boldsymbol{\theta}$ est composé de paramètres de plusieurs niveaux hiérarchiques, la loi *a posteriori* conditionnelle pour un paramètre du niveau n ne dépendra que des niveaux $n-1$ et $n+1$. Nous nous retrouvons donc souvent avec des lois *a posteriori* conditionnelles avec un faible nombre de dépendances qui sont plus faciles à exprimer et échantillonner.

Si pour tout n, les lois *a priori* sont conjuguées par les vraisemblances correspondantes, on parle également d'**échantillonneur de Gibbs** [3, Sect. 11.5]. Si $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_N]$ est le vecteur des paramètres et $\mathbf{Y}$ les données, les lois *a posteriori* conditionnelles s'écrivent $\mathrm{p}(\theta_n \mid \mathbf{Y}, \theta_1, \ldots, \theta_{n-1}, \theta_{n+1}, \ldots, \theta_N)$ pour $n = 1, \ldots, N$. Comme l'échantillonneur de Gibbs est intégré à l'intérieur d'une chaîne de Markov, le conditionnement se fait par les dernières valeurs connues des paramètres. À l'itération (k), l'échantillonnage du paramètre θ_n se fait donc sous la loi

$$\mathrm{p}\left(\theta_n \mid \mathbf{Y}, \theta_1^{(k)}, \ldots, \theta_{n-1}^{(k)}, \theta_{n+1}^{(k-1)}, \ldots, \theta_N^{(k-1)}\right).$$

L'échantillonnage pour les paramètres dont les lois sont conjuguées est facile à faire. Comme la plupart des logiciels permettant de faire du calcul algorithmique disposent

d'échantillonneurs classiques, il suffit de mettre à jour à chaque itération les paramètres des lois *a posteriori* conditionnelles et d'échantillonner sous ces lois. Le cas où il n'y a pas de conjugaison pour une partie des paramètres est étudié dans le paragraphe suivant.

Remarque ⟨2.36⟩ (Convergence) *Sous des conditions générales, la loi échantillonnée à l'aide des échantillonneurs de Metropolis-Hastings et de Gibbs converge vers la loi cible. Nous référons le lecteur à [3, Sect. 11.4] pour plus de détails.*

Gibbs hybride

L'échantillonneur de Gibbs, tel qu'il est présenté dans la sous-section précédente, fonctionne avec des lois conjuguées. Cependant, que faire quand la loi *a posteriori* n'a pas de forme usuelle pour laquelle existent des générateurs d'échantillons, quand la loi *a priori* n'est pas conjuguée par la vraisemblance associée [3, Sect. 11.5] ?

Nous avons introduit précédemment l'algorithme de Metropolis-Hastings qui permet d'échantillonner sous une loi cible. Au lieu d'intégrer une étape d'échantillonnage explicite, nous intégrons pour les paramètres concernés **une étape** de MH (avec la loi de proposition que l'on souhaite) pour proposer des échantillons, c'est-à-dire **un tirage aléatoire** sous la loi cible (et non tout une boucle MH). Nous appelons l'intégration d'une étape MH (ou d'autres moyens d'échantillonner une loi *a posteriori* conditionnelle non usuelle) dans un cycle de Gibbs par le terme **Gibbs hybride**.

Remarque ⟨2.37⟩ (Structure de Gibbs et Metropolis-Hastings) *L'ouvrage [3] indique explicitement dans la Sect. 11.4 (contrairement à beaucoup d'ouvrages sur le sujet d'ailleurs) que la loi de proposition d'un algorithme de Metropolis-Hastings peut dépendre de l'itération courante. Il en donne d'ailleurs un exemple en précisant que le noyau gaussien d'une Marche Aléatoire peut varier suivant les propriétés de convergence de l'échantillonnage actuel, voir Sect. 11.9 de la référence.*

Ainsi, on peut relier la structure de Gibbs à l'échantillonneur de Metropolis-Hastings : chaque itération (k) consiste en N étapes, N étant le nombre de paramètres. À l'étape n, la structure de Gibbs permet d'échantillonner uniquement le candidat θ_n^{p} pour la $n^{\text{ème}}$ entrée du vecteur de paramètres à partir la loi de proposition suivante :

$$\theta_n^{\mathrm{p}} \sim \pi\left(\theta_n \mid \theta_1^{(k)}, \ldots, \theta_{n-1}^{(k)}, \theta_{n+1}^{(k-1)}, \ldots, \theta_N^{(k-1)}\right).$$

La mise à jour s'effectue uniquement au niveau de la $n^{\text{ème}}$ entrée du vecteur.

Si on peut échantillonner la loi a posteriori *conditionnelle grâce à une conjugaison de loi, c'est cette dernière qui fera office de loi de proposition :*

$$\pi\left(\theta_n \mid \theta_1^{(k)}, \ldots, \theta_{n-1}^{(k)}, \theta_{n+1}^{(k-1)}, \ldots, \theta_N^{(k-1)}\right) = \mathrm{p}\left(\theta_n \mid \mathbf{Y}, \theta_1^{(k)}, \ldots, \theta_{n-1}^{(k)}, \theta_{n+1}^{(k-1)}, \ldots, \theta_N^{(k-1)}\right).$$

On effectue un échantillonnage explicite car dans Éqn. (2.40) le rapport se simplifie à 1*, synonyme d'acceptation systématique du nouvel échantillon. La loi de proposition propose donc* le *nouvel échantillon ; voir Sect. 11.5 de la référence en question.*

Dans le cas contraire, on effectue une étape de Metropolis-Hastings comme décrit précédemment.

On a alors le dégradé suivant :

- *Toutes les lois de proposition sont des lois* a posteriori *conditionnelles issues d'une conjugaison. Dans ce cas, on se retrouve avec un échantillonneur de Gibbs.*
- *Certaines lois de proposition sont des lois* a posteriori *conditionnelles issues d'une conjugaison, d'autres ne le sont pas. C'est ce que l'on appelle l'échantillonneur de Gibbs hybride.*
- *Aucune loi de proposition est une loi* a posteriori *conditionnelle issue d'une conjugaison. En procédant étape par étape à l'intérieur de chaque itération, on se retrouve avec un échantillonneur séquentiel qui effectue des étapes de Metropolis-Hastings pour chaque paramètre.*

Mises en garde quant aux lois a priori impropres

Les distributions *a posteriori* obtenues incluant une loi *a priori* impropre sont à traiter avec beaucoup de précaution [48]. En effet, la distribution *a posteriori* peut ne pas être intégrable et n'est donc pas une densité de probabilité ! Cela peut créer des effets numérique inattendus, même si ce n'est pas toujours le cas, comme constatent la référence ci-dessus ainsi que [32]. Avant de commencer l'inférence, la fonction résultante est donc toujours à étudier et son intégrabilité à vérifier. Alors que la démonstration du caractère propre d'une loi peut être facile dans des cas simples, elle est souvent analytiquement infaisable pour des modèles plus compliqués. Certains auteurs ont le réflexe de vouloir approcher la fonction candidate par des calculs numériques, notamment de l'échantillonnage stochastique. Pour la démonstration, ils tirent des échantillons selon une fonction qui n'est peut-être pas une distribution. Malgré des résultats pratiques « satisfaisants », la preuve théorique échoue éventuellement et ne peut donc mener à une conclusion par rapport à l'intégrabilité de la fonction.

Dans [39], les auteurs se trouvent dans la situation suivante : un modèle à effet mixte, incluant un certain nombre de paramètres, est proposé afin de modéliser un phénomène. Toutes les lois *a priori* sont impropres, la vraisemblance est gaussienne. Par des calculs simples, on montre que toutes les lois *a posteriori* conditionnelles sont **propres**. À ce moment, on peut avoir envie de croire que, vu le caractère propre de toutes les conditionnelles, la distribution *a posteriori* totale doit l'être aussi. *Hélas, ce n'est pas le cas* puisqu'ils montrent par des calculs analytiques le caractère impropre de la loi *a posteriori*.

Dans nos évaluations algorithmiques, nous avons également constaté que, lorsque nous utilisions des lois *a priori* impropres pour les paramètres, certaines estimations conjointes des paramètres consistaient en des valeurs aberrantes comme par exemple une concentration protéique de 10^{30} ng/mℓ, *i.e.* une concentration de 1 billiard de tonnes par milli-litre. Cependant, nous n'obtenions pas systématiquement ce résultat aberrant : une réévaluation apportait un résultat plus proche de la réalité. La programmation ne montrait pas de défaut, et l'analyse des chaînes MCMC dans les cas « normaux » ne montrait aucune anomalie. Celles associées aux valeurs aberrantes montraient un saut soudain à des valeurs extrêmes pour les paramètres modélisés par une loi impropre, autorisant *a priori* toute valeur réelle. Nous avons alors substitué les lois impropres par des lois vaguement informatives, *i.e.* des lois normales avec une précision proche de 0 et une loi gamma avec les paramètres de forme α proche de 0 et d'échelle β très grand. L'introduction des lois propres a permis de supprimer les valeurs extrêmes et aberrantes, depuis ce changement toutes les estimations sont « raisonnables ».

2.9.2 Bayésien Variationnel

Les algorithmes d'échantillonnage approchent empiriquement des lois avec des tirages aléatoires, les lois pouvant être déduites par les histogrammes correspondants. Avec un nombre suffisamment grand d'échantillons, l'histogramme converge vers la loi *a posteriori* ciblée. Les **méthodes bayésiennes variationnelles** cependant fournissent une expression analytique de l'approximation de la loi *a posteriori* jointe sous forme de produit de lois *a posteriori* marginales :

$$\mathrm{p}_{\Theta|\mathbf{Y}}(\boldsymbol{\theta} \mid \mathbf{Y}) \approx \tilde{\mathrm{p}}_{\Theta|\mathbf{Y}}(\boldsymbol{\theta} \mid \mathbf{Y}) = \prod_{n=1}^{N} \mathrm{p}_{\Theta_n|\mathbf{Y}}(\theta_n \mid \mathbf{Y}).$$

Cette écriture force une indépendance *a posteriori* des sous-ensemble de paramètres $\theta_1, \ldots, \theta_n$. La meilleure approximation est trouvée en minimisant la divergence de Kull-

back-Leibler entre l'approximation de l'*a posteriori* par lois marginales et la loi *a posteriori* :

$$\hat{\mathrm{p}}(\theta \mid \mathbf{Y}) = \underset{\tilde{\mathrm{p}}(\boldsymbol{\theta}|\mathbf{Y})}{\operatorname{argmin}} \mathrm{KL}\left(\tilde{\mathrm{p}}(\boldsymbol{\theta} \mid \mathbf{Y}) \| \mathrm{p}(\boldsymbol{\theta} \mid \mathbf{Y})\right),$$

où $\hat{\mathrm{p}}$ dénote la distribution estimée. Tout en étant la meilleure approximation au sens de la divergence de Kullback-Leibler, cette distribution n'est pas et ne converge pas vers l'*a posteriori*. Ensuite, les calculs « classiques » peuvent être appliqués à partir de cette distribution : estimateurs divers, approximation de Laplace dans le contexte du choix de modèle (voir [49]), échantillonnage de la loi et intégration avec des algorithmes de type Monte Carlo.

Cependant, nous n'utilisons pas le bayésien variationnel dans cette thèse au profit des approches d'échantillonnage stochastique, conceptuellement mieux adaptées à notre problème car ces dernières nous permettent l'accès aux estimées autres que le maximum et aux interactions entre les paramètres à travers la loi *a posteriori*.

Une introduction aux méthodes bayésiennes variationnelles peut être trouvée dans [50]. Des extensions à une nouvelle approche bayésienne variationnelle sont décrites dans [51, 52], des exemples dans la littérature se trouvent par exemple dans [53–55].

3 Modélisation physique & probabiliste

Le processus d'acquisition d'une donnée en protéomique clinique pour l'estimation de la concentration de P protéines commence par l'extraction du liquide porteur (urine, sang, ...) et se termine par la fourniture d'un signal. Entre ces deux extrémités, l'échantillon biologique subit une cascade de processus, illustrée sur la Fig. 3.1. Cette chaîne analytique n'est pas spécifique à un seul instrument, comme nous le présenterons ci-dessous : le spectromètre de masse peut donner par exemple une sortie en deux dimensions qui balaie toutes les masses (« *Full-Mass-Spectrometry* », « Full-MS ») ou il peut fournir des successions de signaux mono-dimensionnels pour des masses pré-sélectionnées (« *Selected Reaction Monitoring* », « SRM »). Quel que soit l'instrument considéré, la structure en cascade nous permettra de proposer un modèle hiérarchique de l'acquisition [1], [117]. Les modèles physique et probabiliste d'acquisition d'une sortie de la chaîne analytique sont l'objet de ce chapitre.

3.1 Préparation de l'échantillon biologique

Afin d'obtenir des échantillons « injectables » dans l'instrument considéré, les échantillons « bruts » doivent être préparés. C'est dans ces étapes de préparation que l'on trouve beaucoup de sources de variabilité que l'on va schématiser par la suite. Cette variabilité se traduira par un gain sur les protéines, $\boldsymbol{\psi} \in \mathbb{R}_+^{P\times P}$. Il s'agit d'une matrice diagonale qui a comme entrée $[\boldsymbol{\psi}]_{pp} = \psi_p$ la valeur du gain subi par la concentration de la protéine p.

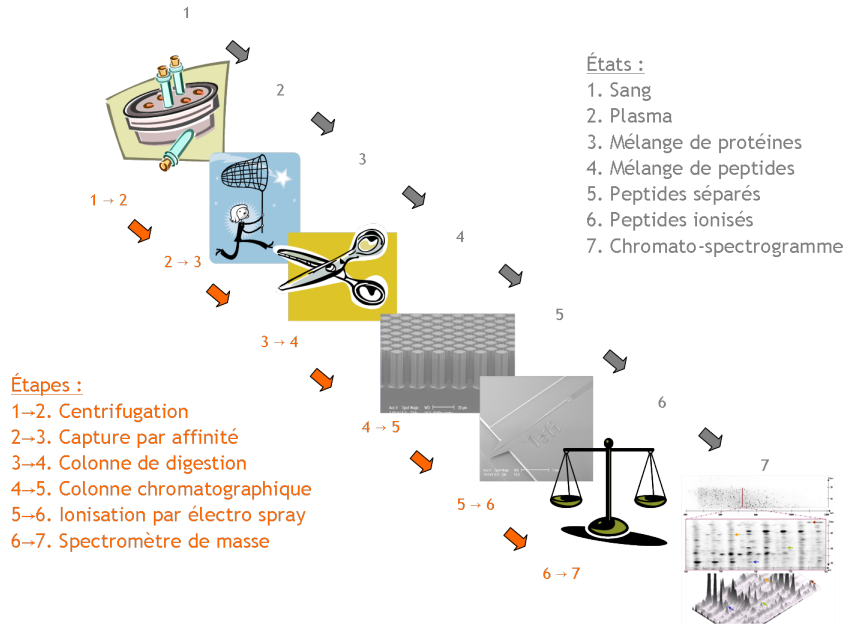
3.1.1 Prélèvement des échantillons

L'échantillon biologique dans cette thèse est prélevé sous forme de sang. (D'autres protocoles prévoient des prélèvements d'autres fluides comme la salive, les muqueuses, l'urine, la transpiration, *et cetera*). Afin de séparer le sérum, contenant notamment les protéines, des globules rouges et des globules blancs, une étape de centrifugation est entreprise. À la fin de celle-ci, l'échantillon de sérum est aliquoté et congelé.

3.1.2 Réduction de la complexité de l'échantillon

D'après les dernières statistiques de l'HUPO (*Human Proteime Organisation*) [56], le plasma humain contient jusqu'au moins trois mille protéines. Parmi cette quantité, il n'y a qu'une infime partie que nous cherchons à étudier, *cf.* Fig. 1.3. Ces quelques protéines cibles ne sont généralement pas majoritaires : rappelons que 97% du contenu du plasma est constitué d'une vingtaine de protéines seulement, les quelques 2980 autres se partageant les trois pour-cent restants ! Mais certaines de ces protéines de basse concentration sont très importantes pour le diagnostic de maladies, comme par exemple l'Amyloid, ou plus précisément son peptide Amyloid bêta, pour la maladie d'Alzheimer [57].

Fig. 3.1 Schéma de la chaîne d'analyse.



3.1.3 Capture par affinité

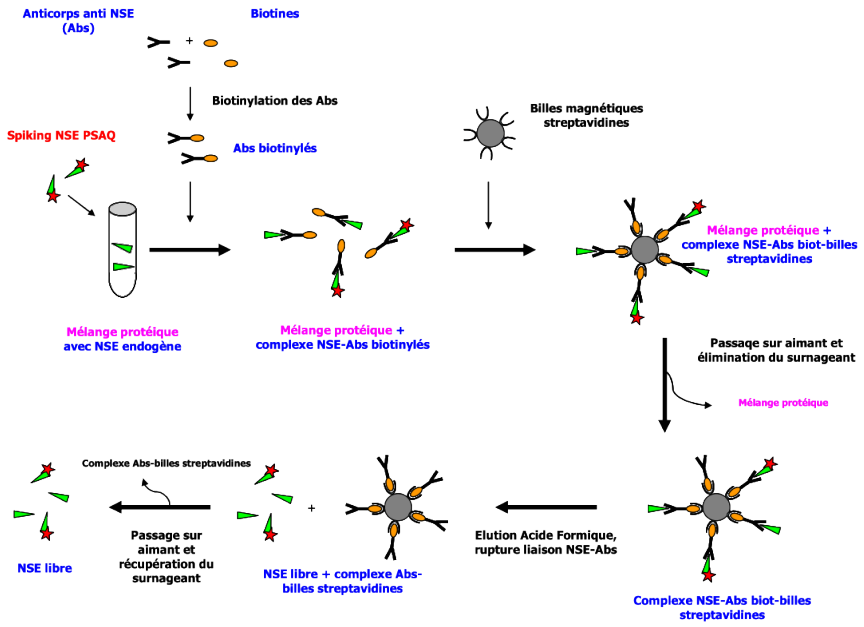
L'étape de capture par affinité[1] consiste à rehausser la concentration des protéines cibles par rapport aux autres protéines. Son fonctionnement est schématisé dans la Fig. 3.2. Pour cela, des anticorps biotinylés spécifiques sont ajoutés au mélange. Uniquement les protéines cibles se collent à ces anticorps, les autres ne trouvant pas de partenaire. Ces couples sont ensuite capturés par des billes magnétiques auxquelles les biotines réagissent : elles sont attirées. Dans le mélange se trouvent alors des protéines cibles accrochées à des billes magnétiques et d'autres protéines non cibles. En passant sur un aimant, les billes magnétiques sont fixées, les autres éléments sont libres. Nous pouvons alors laver le mélange ce qui permet d'éliminer les autres protéines. Notre mélange est ainsi purifié et ne contient plus que les protéines cibles accrochées aux billes. Les protéines sont ensuite séparées de leurs anticorps, puis un deuxième passage sur un aimant élimine les billes. Ainsi, le mélange est décomplexifié et est censé ne plus contenir que les protéines cibles.

3.1.4 Marquage isotopique

On peut s'intéresser à une protéine qualitativement ou quantitativement. Dans le premier cas, on pose la question : est-ce qu'une protéine donnée est présente ou absente sous des conditions spécifiées ? Ces méthodes permettent par exemple l'identification de protéines. Dans le cas que nous proposons de regarder, nous voulons connaître la quantité des protéines, leur concentration. La réponse serait simple si le système était stable d'une expérience à l'autre. Hélas, les variations sont grandes. Pour contourner ce problème, il existe des méthodes de calibrage interne. Les méthodes de type « marquage isotopique » consistent à marquer les protéines d'intérêt à l'aide de molécules alourdies et d'analyser conjointement le mélange « marqué + natif ». Par rapport aux molécules natives, certains atomes comme l'azote, le carbone ou l'oxygène sont substitués par des atomes isoto-

1. Cette étape est utilisée dans la chaîne d'analyse de nos collègues de l'Institut de Recherches en Technologies et Sciences pour le Vivant au CEA Grenoble.

Fig. 3.2 Stratégie d'immuno-capture de la protéine NSE [58].



piquement alourdis dans les molécules marquées. Elles ont alors les mêmes propriétés physico-chimique, mais une masse différente. Le *standard*, *i.e.* l'ensemble des molécules marquées, subit donc de manière analogue les étapes de la chaîne d'analyse comme la digestion (si ajouté auparavant dans le mélange) ou la séparation chromatographique, mais elles sont séparées lors de l'analyse de masse. Cela veut dire que nous retrouvons pour le couple marqué/natif deux mêmes signaux chromatographiques (même forme, même position sur l'axe chromatographique) à des amplitudes correspondant à leur concentrations respectives, mais à des positions sur l'axe masse différentes, comme l'illustre la Fig. 3.7 (page 50) pour des données « Full-MS » et la Fig. 3.3 (page 46) pour des données « SRM ». Ainsi, on peut relier l'intensité du signal de la molécule marquée à sa concentration connue et l'intensité du signal de la molécule native à sa concentration inconnue.

Parmi les méthodes de marquage isotopique utilisées fréquemment en protéomique, nous allons nous concentrer sur deux : le marquage PSAQ au niveau de la protéine, et le marquage AQUA au niveau du peptide. La différence est au niveau de la technologie de synthèse. Les peptides peuvent être synthétisés atome par atome, ce qui n'est pas le cas des protéines qui sont produites par des mécanismes cellulaires plus complexes. D'autres standards sont présentés dans [59].

Protein Standard Absolute Quantification (PSAQ)

Pour qu'une molécule marquée soit efficace par rapport à la quantification protéique, elle doit subir le plus d'étapes possible de la chaîne d'analyse conjointement à la molécule cible. Autrement dit, plus le standard est ajouté tôt, plus les rendements et gains peuvent être pris en compte lors de la quantification. La méthode de marquage PSAQ (*Protein Standard Absolute Quantification*) [59] propose l'adjonction de protéines standards à concentration x^* connue dans l'échantillon biologique. Ainsi, les PSAQ subissent toutes les modifications, pertes, tous les gains, rendements et passent par toute la chaîne analytique. Nous pouvons déduire directement des informations sur la concentration des protéines.

Fig. 3.3 Stratégie d'analyse quantitative de peptide. On ajoute aux molécules natives (« No labeling ! ») le standard interne (« Internal standards »), ici des peptides marqués isotopiquement. En sortie, on obtient un signal pour les molécules natives, et un deuxième signal pour les molécules standards. Figure issue de [60].

Absolute Quantification (AQUA)

Une autre méthode, moins chère, est la fabrication de peptides alourdis. En effet, ceux-ci peuvent être rajoutés à l'échantillon biologique relativement tôt, mais ne calibrent les gains et rendement qu'à partir du niveau peptides. Donc, nous ne pouvons surveiller de manière interne le gain entre protéines et peptides.

Comment accéder alors aux gains au-dessus des peptides ? Pour cela, nous supposons qu'entre les acquisitions issues d'une même préparation, ces gains ne fluctuent pas. Ensuite, les données acquises à un jour donné sont accompagnées de données dites de « Contrôle de Qualité ». Il s'agit d'un échantillon similaire à ceux que l'on veut analyser, mais avec une concentration native connue. C'est de celui-ci dont nous nous servirons pour calibrer les gains de préparation et de digestion.

3.1.5 Colonne de digestion

Les protéines étant des molécules trop complexes pour être directement analysées en spectrométrie de masse, celles-ci sont classiquement fragmentées en peptides par un processus de digestion préalable. La digestion par trypsine est réalisée après les acides aminés R (arginine) ou K (lysine) et produit donc des peptides aux séquences en acides aminés bien déterminées.

Il est possible qu'une protéine donne plusieurs fois le même peptide, et plusieurs protéines peuvent produire un peptide identique. On parle dans ce cas de *peptides partagés*. Si un peptide est unique à une protéine donnée dans le protéome considéré, on parle de *peptide protéotypique*.

Soit $d_{ip} \in \mathbb{N}_0$ le nombre de peptides i générés par une protéine indicée par p. Ainsi, la quantité du peptide i issu de toutes les protéines cibles $p = 1, \ldots, P$ est donnée par

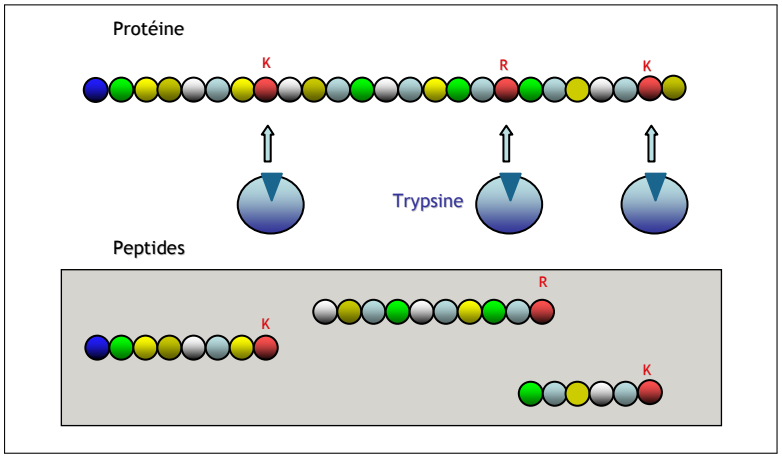
$$\kappa_i = \sum_{p=1}^{P} d_{ip}\, \psi_p x_p \tag{3.1}$$

la somme des concentrations des protéines x_p pondérées par le gain de préparation ψ_p et le gain de digestion d_{ip}.

Nous pouvons écrire cette relation d'une manière matricielle. Pour cela, la matrice $\mathbf{D} = \left(d_{ip}\right)_{i,p} \in \mathbb{N}_0^{I \times P}$ avec $i = 1, \ldots, I$, $p = 1, \ldots, P$ est la matrice des gains de digestion. Ainsi, le vecteur des quantités de peptides $\boldsymbol{\kappa}$ est donné par

$$\boldsymbol{\kappa} = \mathbf{D}\boldsymbol{\psi}\boldsymbol{x} \tag{3.2}$$

Fig. 3.4 Principe de la digestion par la trypsine. Chaque cercle représente un acide aminé. La trypsine coupe après la lysine et l'arginine. La proportion de coupes effectuées par coupes possibles est représentative de l'efficacité de la digestion qui dépend de plusieurs facteurs. Figure fournie par le *Laboratoire d'Étude de la Dynamique des Protéomes*, iRTSV, CEA Grenoble dans le cadre de [1].



correspondant au produit de la matrice de digestion $\mathbf{D}$ avec la concentration des protéines cibles $\boldsymbol{x}$ après application du gain $\boldsymbol{\psi}$. Cette modélisation permet de passer de la quantité des protéines en entrée de la chaîne d'analyse à la quantité des peptides à la fin des étapes de préparation.

La digestion peut avoir deux facteurs de bruit : un bruit additif, noté ε_κ, et un bruit multiplicatif. Ce dernier, noté $\chi_i \in \mathbb{R}_+$, traduit l'efficacité de la digestion pour le peptide i. On regroupe ces *rendements de digestion* dans la matrice diagonale $\boldsymbol{\chi} \in [0,\ 1]^{I \times I} = \mathrm{diag}([\chi_1, \ldots, \chi_I])$.

Enfin, le modèle retenu pour la digestion est

$$\boldsymbol{\kappa} = (\boldsymbol{\chi}\mathbf{D}\boldsymbol{\psi})\boldsymbol{x} + \boldsymbol{\varepsilon}_\kappa \tag{3.3}$$

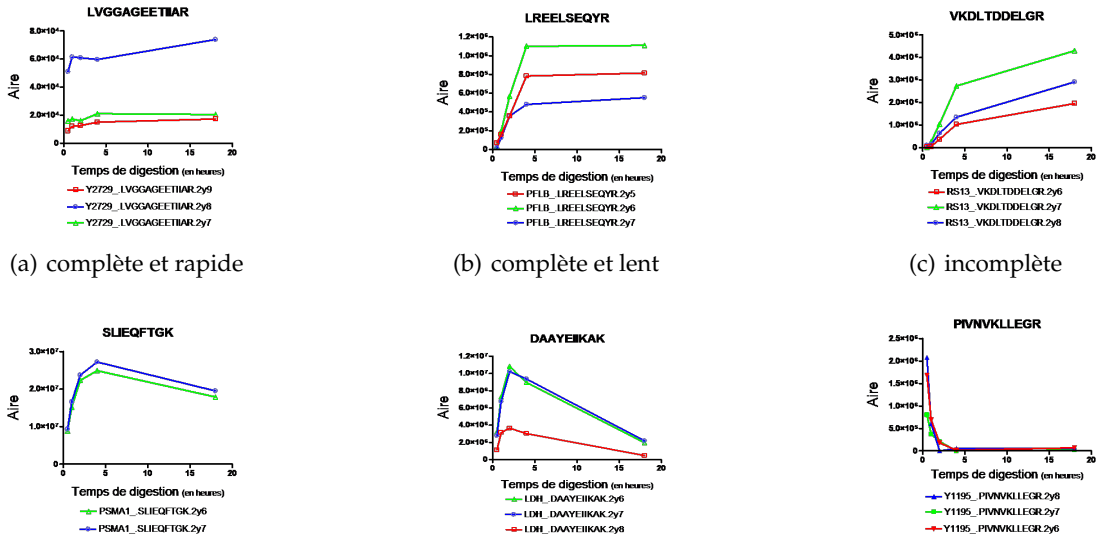
Détaillons l'aspect « rendement de digestion ». Le processus de digestion est sujet à une cinétique de réaction [61] : la digestion n'est pas instantanée, il faut du temps pour que la quantité de trypsine injectée découpe les protéines. Cependant, elle n'est pas la même pour tous les peptides, et n'a pas les mêmes comportements asymptotiques : il y a des cinétiques complètes et rapides (*i.e.* la digestion est considérée complète après très peu de temps), des cinétiques complètes mais lentes, des cinétiques incomplètes. D'autre part, la stabilité des peptides obtenus dans la solution est un facteur ayant une conséquence sur la quantité des peptides à l'issue de l'étape de digestion. Cette stabilité peut être perturbée par plusieurs raisons (par exemple une interaction avec d'autres molécules) qui ne dépendent *strictu sensu* du phénomène enzymatique de digestion de protéines en peptides. Ainsi, sur la Fig. 3.5, [62], on observe pour différents temps de digestion l'aire du signal SRM pour plusieurs peptides.

La cinétique varie non seulement suivant le temps de digestion, mais également en fonction de la température ϑ du mélange, et en fonction du pH du solvant.

Pour faciliter la compréhension, nous nous servons de l'exemple simple suivant.

Exemple ⟨3.1⟩ *Considérons le cas simple de la digestion d'une protéine, ①②, en deux peptides, peptide ① et peptide ②. Pour plus de simplicité, le gain de préparation vaut $\psi = 1$. La digestion est censée séparer les deux peptides ; autrement dit, ①② $\longrightarrow$ ① + ②.*

Fig. 3.5 Cinétiques de digestion observées en suivant l'évolution de l'aire sous la courbe chromatographique. Figures issues de [62].



(a) complète et rapide (b) complète et lent (c) incomplète

(d) en cloche lente (e) en cloche rapide (f) disparition complète

Nous pouvons donc modéliser la digestion comme le résultat d'une partie « digestion complète » plus une partie « digestion incomplète » ; en symboles ①② $\longrightarrow \chi_1$① $+ \chi_2$② $+ \chi_{12}$①② *avec* $\chi_i = \chi_i(t_{dig}, \vartheta, pH) \in [0,\ 1]$ *pour* $i \in \{1, 2, 12\}$.

Les molécules ① *et* ② *sont appelées molécules* monopeptidiques *tandis que* ①② *est une molécule* polypeptidique. *Une propriété importante est celle de la conservation de la quantité : le nombre de peptides avant la digestion (i.e. non décomposé dans les protéines) est égal au nombre de peptides après l'étape de digestion. Dans l'exemple ci-dessus, nous avons alors que* $\chi_1 = \chi_2 = 1 - \chi_{12}$. *Cette relation se complique dramatiquement dès que nous considérons une protéine à trois peptides ou plus. Dans les expériences utilisant le mode SRM, nous ne nous intéressons classiquement qu'aux monopeptides.*

Suite à ces observations, nous définissons le nombre de peptides I *comme la somme du nombre des monopeptides* I^{mono} *et des polypeptides* I^{poly}. *Dans l'exemple ci-dessus, nous avons* $I = 2 + 1 = 3$.

Pour la modélisation de la digestion, chaque polypeptide doit être décrit dans la matrice $\mathbf{D}$. *Ainsi, le vecteur des quantités peptidiques contient également* $I^{mono} + I^{poly}$ *entrées.*

Finalement, nous écrirons les rendements de digestion χ_i *pour* $i = 1, \ldots, I^{mono} + I^{poly}$ *dans un vecteur* $\chi \in [0,\ 1]^I$. *Pour reprendre l'exemple précédent,* $\chi = [\chi_1, \chi_2, \chi_{12}]^{\mathrm{T}}$.

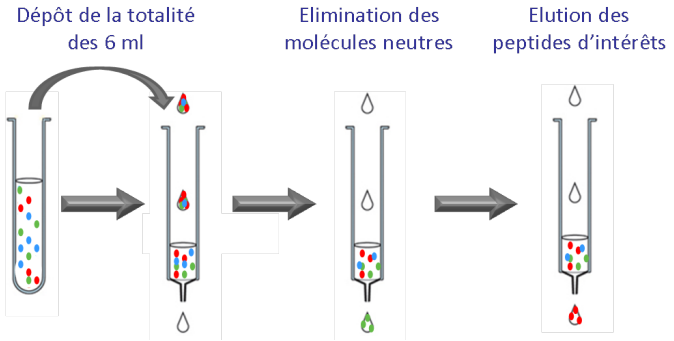
Remarque ⟨**3.2**⟩ *Les rendements de digestion peuvent être calculés en utilisant la base de données PeptideCutter [107], [63, 64].*

3.1.6 Fractionnement peptidique

Le but du *fractionnement* est de dessaler et concentrer les peptides ce qui réduit la complexité de l'échantillon. Cette étape qui fait partie du protocole expérimental de l'acquisition des données SRM a l'intérêt d'être peu coûteuse et peu risquée par rapport à une purification de protéines où le risque de co-élimination avec l'albumine subsiste.

Afin de fractionner, l'échantillon est injecté dans un tube contenant une surface en résine. Une solution à un pH précis est ajoutée ensuite pour la réalisation de la première fraction. Après élution complète de cette solution, l'échantillon est fractionné une

Fig. 3.6 Stratégie du fractionnement peptidique. Illustration issue de [62].



deuxième fois avec une solution à un pH différent, et il en est de même pour réaliser une troisième fraction. Ce fractionnement peut se faire suivant un protocole automatisé (10 minutes) ou manuel (45 minutes).

3.2 Chromatographie liquide

Le but d'une colonne de chromatographie[xvi] est de séparer les composantes d'un mélange suivant différentes propriétés d'affinité. Dans notre application, nous séparons les peptides. Les peptides sont ensuite caractérisés par leur *temps d'élution* ou *temps de rétention*, noté τ_i pour le peptide i : toutes les molécules ayant le même temps de rétention ont la même propriété chromatographique et sont donc du même type.

Le débit moléculaire $c_i(t)$ d'un peptide i en entrée de la colonne de chromatographie se modélise donc *idéalement* sous forme d'un Dirac[xvii] : $c_i(t) \propto \delta(t - \tau_i)$ où τ_i est le temps d'élution du peptide. Cependant, en sortie on peut voir une déformation de ce Dirac dû à la réponse de l'instrument, voir par exemple Fig. 3.7. Nous l'approchons par une gaussienne centrée en τ_i, sa largeur étant déterminée par son inverse variance notée λ_i [1, Sect. 2.2] :

$$\mathscr{C}_i(t; \tau_i, \lambda_i) = \mathcal{N}(t; \tau_i, \lambda_i). \tag{3.4}$$

Remarque ⟨3.3⟩ (Discrétisation) *Soit $T^e_{\mathscr{C}}$ le temps d'échantillonnage du signal chromatographique. Nous notons $\mathscr{C}_i[n] = \mathscr{C}_i(nT^e_{\mathscr{C}}; \tau_i, \lambda_i)$ le profil chromatographique $\mathscr{C}_i$ du peptide i échantillonné au point n.*

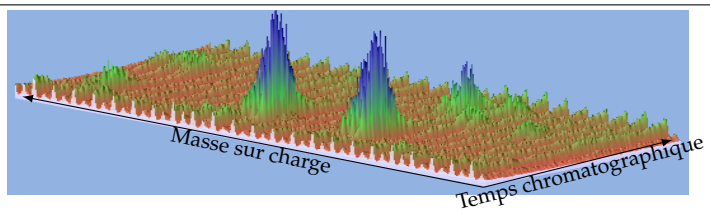
Remarque ⟨3.4⟩ (Justification de la Gaussienne) *La littérature concernant la forme d'un pic chromatographique est vaste et le choix du modèle adéquat est une tâche très difficile. Nous avons choisi d'approcher la forme par une gaussienne, même si ce choix ne limite pas la méthode que nous présenterons dans le suite. En effet, d'autres formes peuvent être utilisées quasiment* ad hoc.

Les causes de l'asymétrie, de queues lourdes ou d'autres phénomènes déformant le pic sont variées puisqu'il s'agit d'une représentation du temps d'arrivée d'un groupe de molécules. Ces dernières ne voyagent pas à la même vitesse et n'ont pas parcouru précisément la même distance dans la colonne chromatographique. Les ouvrages [65, 66] proposent l'utilisation d'équations différentielles pour décrire la propagation des molécules dans une colonne de chromatographie liquide.

L'utilisation de modèles inadaptés peut avoir deux conséquences majeures : une détermination erronée de la position du pic chromatographique et de la quantité. La référence [67] montre ces deux effets — certes, modérés mais présents — en comparant des formes gaussiennes.

Il est important de voir que les mesures (de quantité, de position, ...) que nous entreprenons se font conjointement à une molécule calibrante. Ainsi, les deux conséquences sont atténuées,

Fig. 3.7 Visualisation d'un extrait de spectre LC-MS [1] où on voit bien le caractère gaussien du profil chromatographique.



comme par exemple une mauvaise évaluation de la position du pic. En effet, en sélectionnant une mauvaise position on commet propotionnellement « la même erreur » pour le pic de la molécule native et celui de la marquée si la forme des deux réponses est la même, et le rapport reste le même.

Ensuite, la forme gaussienne peut être facilement et rapidement mise en place dans nos développements. Elle nous permet d'accélérer certains calculs par des fonctions proposées dans les boîtes à outils. Les inconvénients sont petits et les erreurs de quantification négligeables devant le gain calculatoire, comme le montre [118].

Remarque ⟨3.5⟩ *L'instrument utilisé pour les données LC-Full-MS fonctionne en mode gradient. Dans ce mode, chaque pic chromatographique d'un peptide a la même largeur $\lambda_{\mathscr{C}}$. Dans d'autres modes, ce n'est pas forcément le cas, voir aussi [1, Sect. 2.2].*

La largeur des pics est réputée stable d'une expérience à l'autre. Nous trouvons un comportement contraire concernant les temps de rétention : des variations relativement fortes, de l'ordre d'une minute, peuvent être observées. En effet, des phénomènes de saturation de la colonne peuvent déformer les pics et entraîner un décalage sur le temps de rétention moyen.

3.3 Ionisation par électro-nébulisation

Le spectromètre de masse utilisé transforme un courant électrique issu d'un gaz d'ions en signal. Or, les peptides à la sortie de la colonne de chromatographie sont en phase liquide. Pour produire les ions qui seront injectés dans le spectromètre de masse, plusieurs méthodes existent et d'autres sont toujours en élaboration. Celle qui est utilisée ici pour la fourniture des données est l'électro-nébulisation. Sa modélisation est décrite dans [1, Sect. 2.3].

Le principe est de former et de charger de fines gouttelettes à partir du liquide, sous l'action d'un champ électrique appliqué entre l'aiguille de sortie de la colonne de chromatographie et le spectromètre de masse. Le liquide s'évapore alors, libérant les molécules qui sont injectées ensuite dans le spectromètre.

3.4 Spectrométrie de masse

Un spectromètre de masse est un appareil qui prend en entrée une certaine quantité de matière et fournit en sortie un spectre donnant la quantité de molécules en fonction de leur rapport masse sur charge. Le spectromètre agit en deux temps : stockage et lecture. Nous nous intéresserons davantage à la deuxième phase. La modélisation d'un spectromètre enregistrant le temps de vol d'ions est décrite dans [1, Sect. 2.4].

Ce qui suit dans cette section exposera le fonctionnement et le modèle des deux spectromètres de masse considérés dans cette thèse : un spectromètre type « Full-MS » [118], puis un spectromètre type « SRM » [117].

3.4.1 « Full-Mass-Spectrometry »

Pour devenir un ion, un peptide i peut être chargé une ou plusieurs fois. Le nombre de charges j modifie le temps de vol du peptide : un peptide j fois chargé est éjecté j fois plus vite. Ainsi, si le temps de vol du peptide chargé une fois correspond à la masse de m Da, le temps de vol d'un même peptide chargé j fois se situe à la position m/j Da. Ainsi, nous avons un mélange de fonctions de pic aux positions m/j Da dont l'intensité est donnée par la proportion π_{ij} pour le peptide i de porter j charges.

La masse d'un ion dépend notamment de son noyau atomique. Le spectromètre de masse étant assez précis et sensible pour détecter des différences de masse très faibles, nous pouvons distinguer sur le signal deux isotopes d'un même ion précurseur. Cette différence de masse est induite par le nombre de neutrons supplémentaires portés. La masse d'un neutron vaut à peu près 1 Da. Ainsi, idéalement le rapport masse sur charge détecté est un mélange de fonctions de pic aux positions correspondant aux masses. La hauteur de chaque pic est donnée par la proportion π'_{ijk} qu'un peptide i à j charges ait k neutrons supplémentaires. Cet ensemble de pics représentant un même peptide est appelé massif isotopique.

Le nombre de charges intervient de manière inversement proportionnelle dans la transformation
masse $\leftrightarrow$ temps de vol [1, Par. 2.4.3.2]. Les massifs des ions avec j charges se trouvent à $1/j$ fois la masse du peptide considéré.

Les temps de vol d'ions du même type diffèrent légèrement malgré les mêmes conditions de départ. Ceci est dû aux champs électriques parasites engendrés par les ions même qui créent des perturbations, mais évidemment aussi aux pertes par neutralisation. Considérant ce processus comme aléatoire, nous attribuons au temps de vol réel t^{ζ}_{vol} de l'ion ζ une distribution normale centrée autour du temps de vol théorique T^{ζ}_{vol} et avec une largeur $\lambda_{\mathscr{S}}$. Cette dernière peut dépendre de l'ion, mais les observations dont nous disposons montrent que l'instrument utilisé est relativement stable par rapport à cette variable. Nous la considérons indépendante de l'ion mesuré.

La densité de probabilité h_i du temps de vol t^{i}_{vol} d'un peptide i se décompose finalement en un mélange de gaussiennes suivant les masses et les charges possibles du peptide. En identifiant l'ion ζ avec le triplet $(i, j, k) \widehat{=} (\text{peptide}, \text{charges}, \text{neutrons})$, nous écrivons

$$h_i(t^{i}_{\text{vol}}) = \sum_{j=1}^{J} \sum_{k=0}^{K} \pi_{ij} \pi'_{ijk} \mathcal{N}(t^{i}_{\text{vol}}; T^{ijk}_{\text{vol}}, \lambda_{\mathscr{S}}). \tag{3.5}$$

Ayant un nombre suffisant d'événements, nous pouvons modéliser le profil spectrométrique comme étant proportionnel au mélange de distributions normales que nous venons de déduire. Nous définissons $\mathscr{S}_{ijk}(t) = \mathcal{N}(t; T^{ijk}_{\text{vol}}, \lambda_{\mathscr{S}})$.

On note $T^{\text{e}}_{\mathscr{S}}$ le temps d'échantillonnage du signal spectrométrique et $T^{\text{e}}_{\mathscr{C}}$ le temps d'échantillonnage du flux des peptides. Pour exploiter efficacement le signal, on l'organise sous forme d'image, avec un axe représentant les informations spectrométriques, l'autre axe représentant l'information chromatographique [1, Sect. 2.5.2]. Plus précisément, nous changeons le temps d'échantillonnage spectrométrique par la période d'échantillonnage en masse $M^{\text{e}}_{\mathscr{S}}$ puisqu'il y a dualité entre le temps de vol et le rapport masse sur charge, comme nous l'avons évoqué précédemment.

Remarque ⟨3.6⟩ (Notations) *On note $m^0_{\mathscr{S}}$ définissant la première masse échantillonnée. Nous écrivons $h_i[m] = h_i(m^0_{\mathscr{S}} + mM^e_{\mathscr{S}})$ la fonction h_i discrétisée à l'indice de masse m. De la même manière, la discrétisation du profil spectrométrique s'écrit $\mathscr{S}_{ijk}[m] = \mathscr{S}_{ijk}(m^0_{\mathscr{S}} + mM^e_{\mathscr{S}})$.*

Avec les notations précédentes, la sortie *modèle* du système pour le peptide i, l'indice de masse m et l'indice de temps de rétention chromatographique n devient

$$M_i[m,n] = \kappa_i \xi_i \mathscr{C}_i[n] h_i[m], \tag{3.6}$$

où la variable ξ_i incorpore tous les gains que le peptide i a subi, représentant ainsi le produit des gains de préparation, de digestion et d'ionisation. Il est appelé *gain du système*. La matrice $\mathbf{M}_i = (M_i[m,n])_{m,n}$ avec $m = 1, \ldots, N_{\mathscr{S}}$ et $n = 1, \ldots, N_{\mathscr{C}}$ est la matrice modèle du flux du peptide i.

C'est en effet cette dernière matrice qui modélise nos observations et qui correspond au signal modèle chromato-spectrométrique non bruité sous forme d'image. Les mesures modèle non bruitées ont alors la forme suivante :

$$\begin{aligned}
\mathbf{M}[m,n] &= \sum_{i=1}^{I} M_i[m,n] \\
&= \sum_{i=1}^{I} \kappa_i \xi_i \mathscr{C}_i[n] h_i[m] \\
&= \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=0}^{K} \kappa_i \xi_i \pi_{ij} \pi'_{ijk} \mathscr{S}_{ijk}[m] \mathscr{C}_i[n] \\
&= \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=0}^{K} \kappa_i \xi_i \pi_{ij} \pi'_{ijk} \mathscr{S}_{ijk} \mathscr{C}_i^{\mathrm{T}} \right) [m,n]
\end{aligned} \tag{3.7}$$

où $\mathscr{S}_{ijk} = \mathscr{S}_{ijk}[1 : N_{\mathscr{S}}]$ est le vecteur du profil spectrométrique de l'ion (i,j,k) et $\mathscr{C}_i = \mathscr{C}_i[1 : N_{\mathscr{C}}]$ le vecteur du profil chromatographique du peptide i.

Sous forme matricielle, nous pouvons écrire

$$\mathbf{M} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=0}^{K} \kappa_i \xi_i \pi_{ij} \pi'_{ijk} \mathscr{S}_{ijk} \mathscr{C}_i^{\mathrm{T}}. \tag{3.8}$$

Chaque point de mesure $\mathbf{M}[m,n]$ est perturbé par un bruit $\mathbf{B}[m,n]$. Ce bruit peut avoir plusieurs causes. Il peut provenir d'une erreur instrumentale, d'une imperfection de la mesure, mais aussi de la modélisation. Suite aux connaissances que nous avons sur ce bruit, nous choisissons un bruit blanc gaussien centré de moyenne nulle et d'inverse variance γ : $\mathbf{B}[m,n] \sim \mathcal{N}(\mathbf{B}[m,n]; 0, \gamma)$.

L'observation $\mathbf{Y}[m,n]$ s'écrit donc comme somme de la sortie modèle $\mathbf{M}[m,n]$ et d'un terme de bruit $\mathbf{B}[m,n]$. Nous concluons avec l'écriture matricielle de l'observation qui est donnée par

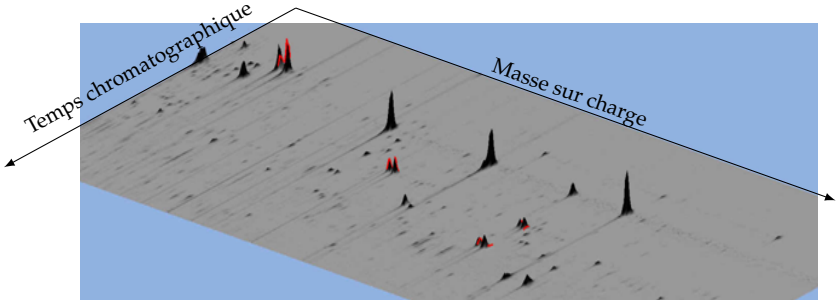
$$\mathbf{Y} = \mathbf{M} + \mathbf{B}. \tag{3.9}$$

Modèle d'une sortie LC-Full-MS avec marquage PSAQ

Dans les développements qui suivent pour cet instrument, nous utilisons la méthode PSAQ. L'échantillon à étudier est alors constitué de l'échantillon biologique enrichi du standard PSAQ. Les instruments fournissent alors un signal joint qui contient le signal du marqué et le signal du natif : l'observation finale $\mathbf{Y}$ peut être écrite comme la somme du signal du natif, $\mathbf{M}$, et du signal du marqué, $\mathbf{M}^*$, perturbé par un bruit blanc gaussien $\mathbf{B}$ centré de moyenne nulle et d'inverse variance γ :

$$\mathbf{Y} = \mathbf{M} + \mathbf{M}^* + \mathbf{B}. \tag{3.10}$$

Fig. 3.8 Réalisation d'un spectre LC-MS de la protéine NSE sur un spectromètre de masse de type trappe ionique du laboratoire EDyP de l'iRTSV/CEA Grenoble. Les peptides issus de la protéine d'intérêt sont entourés en rouge ; les autres pics sont des contaminants. Figure réalisée avec le logiciel MSight [108] de l'Institut Suisse de Bioinformatique.



Identifions les variables dans l'expression de la sortie modèle qui sont affectées par le marquage. Les protéines marquées et les protéines natives ont les mêmes propriétés physico-chimiques. Autrement dit, dans la colonne de chromatographie, elles se comportent de la même manière. Le marquage a été introduit afin d'estimer le gain de système, le gain du peptide i d'une protéine native et celui du peptide i d'une protéine marquée étant le même. Les protéines marquées isotopiquement diffèrent par leur nombre de neutrons. Ainsi, les seules variables impactées par le marquage sont finalement la proportion de neutrons supplémentaires $\pi'_{ijk} \neq \pi'^{*}_{ijk}$ et la masse des molécules observées. Ceci change le profil spectrométrique $\mathscr{S}_{ijk} \neq \mathscr{S}^{*}_{ijk}$:

$$\begin{aligned}\mathbf{Y} &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} \kappa_i \xi_i \pi_{ij} \pi'_{ijk} \mathscr{S}_{ijk} \mathscr{C}_i^{\mathrm{T}} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} \kappa_i^{*} \xi_i \pi_{ij} \pi'^{*}_{ijk} \mathscr{S}^{*}_{ijk} \mathscr{C}_i^{\mathrm{T}} + \mathbf{B} \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} \xi_i \pi_{ij} \left(\kappa_i \pi'_{ijk} \mathscr{S}_{ijk} + \kappa_i^{*} \pi'^{*}_{ijk} \mathscr{S}^{*}_{ijk} \right) \mathscr{C}_i^{\mathrm{T}} + \mathbf{B}. \end{aligned} \tag{3.11}$$

Nous avons alors déduit un modèle mathématique pour l'observation *via* un spectromètre de type Full-MS. On peut en visualiser des réalisations dans la Fig. 3.7 et dans la Fig. 3.8.

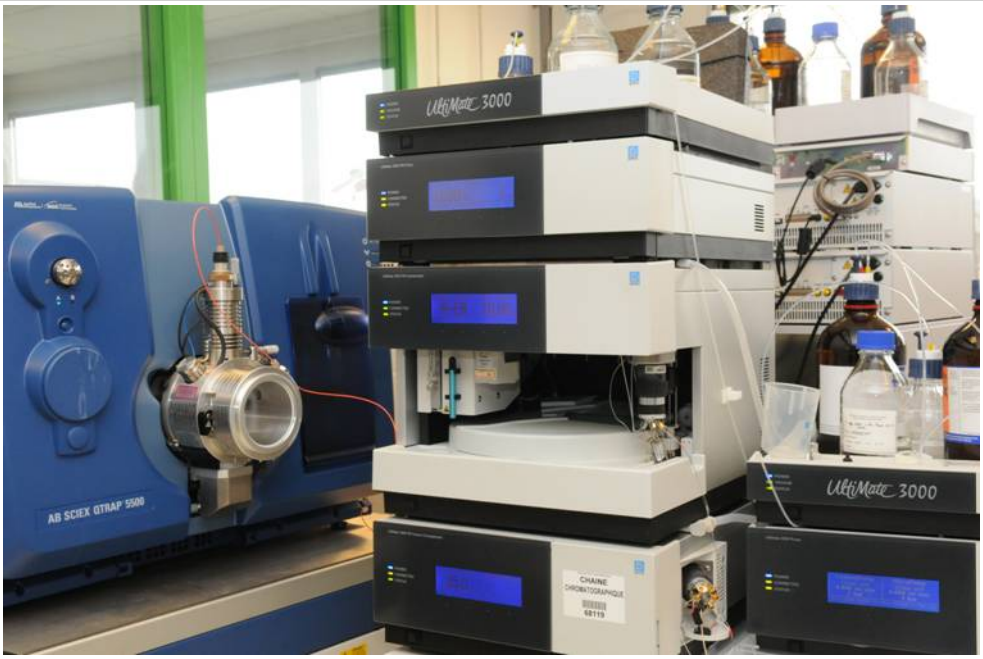
Réécriture de la sortie LC-MS

À l'instar de [1, Sect. 4.1], nous voulons écrire la sortie $\mathbf{Y}$ comme fonction linéaire des paramètres $\boldsymbol{\kappa}$, $\boldsymbol{\kappa}^{*}$ et $\boldsymbol{\xi}$. Pour ce faire, vectorisons la mesure $\mathbf{Y} \in \mathbb{R}^{N_{\mathscr{C}} \times N_{\mathscr{S}}}$ pour obtenir un vecteur $\boldsymbol{y} \in \mathbb{R}^{N_{\mathscr{S}} \cdot N_{\mathscr{C}}}$. Ceci est fait en concaténant les colonnes les unes aux autres. Ensuite, en introduisant des matrices $\mathbf{H}$, $\mathbf{H}^{*}$ et $\mathbf{G}$, éléments de $\mathbb{R}^{N_{\mathscr{S}} N_{\mathscr{C}} \times I}$, nous pouvons écrire

$$\boldsymbol{y} = \mathbf{H}\boldsymbol{\kappa} + \mathbf{H}^{*}\boldsymbol{\kappa}^{*} + \mathbf{B} = \mathbf{G}\boldsymbol{\xi} + \mathbf{B}. \tag{3.12}$$

Ainsi, la sortie $\boldsymbol{y}$ peut être vue comme somme de deux convolutions linéaires des objets $\boldsymbol{\kappa}$ et $\boldsymbol{\kappa}^{*}$, par les matrices réponses $\mathbf{H}$ et $\mathbf{H}^{*}$ respectivement, ou bien comme convolution linéaire de l'objet $\boldsymbol{\xi}$ par la matrice réponse $\mathbf{G}$, perturbée par un bruit $\mathbf{B}$. Pour la déduction et les calculs de ces matrices, nous référons le lecteur à l'Ann. A.6.

Fig. 3.9 Plateforme protéomtique chez l'entreprise bioMérieux à Marcy-l'Étoile : à droite un chromatographe liquide (Dionex, Ultimate 3000), à gauche un spectromètre de masse de type triple quadrupole travaillant en mode SRM (ABSciex 5000QT)



3.4.2 « Selected Reaction Monitoring »

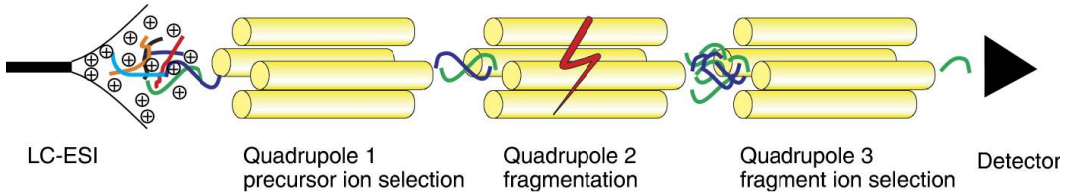
Alors que le Full-MS est un mode de balayage qui sélectionne les données lors du traitement, le mode « Selected Reaction Monitoring » (SRM) cible des masses sélectionnées, ce choix définissant les molécules à filtrer et à fragmenter. L'instrument utilisé pour le mode SRM est constitué d'une succession de trois quadrupoles dont le fonctionnement est le suivant. L'ion précurseur d'un peptide i est isolé et sélectionné dans le premier (Q1) selon sa masse. Ensuite, il entre dans la chambre de collision (Q2) qui est remplie de gaz d'azote. La collision avec les molécules de gaz fait que l'ion est fragmenté. Chaque fragment l associé à l'ion i subit alors un *gain de fragmentation* β_l. Le mélange de fragments est ensuite envoyé dans le troisième quadrupole (Q3) qui est, comme Q1, un filtre de masse : il ne laisse passer que les fragments ayant une masse prédéfinie. Ce sont eux qui arrivent au détecteur et qui, de par la décharge électrique, fournissent le signal – appelé *trace* – de la *transition* ion précurseur/fragment. Dans le contexte « modèle », nous désignerons les transitions par le couple des indices (i, l) ; dans le contexte « protéomique », les transitions sont indiquées par le couple des masses de l'ion précurseur et du fragment associé. La Fig. 3.10 résume et illustre ce mode.

Par la nature du mode SRM, les L traces sont « peptidotypiques », *i.e.* chaque trace est censée ne contenir de l'information que d'un ion seul précurseur issu d'un peptide donné. Autrement dit, chaque indice de fragment l est associé à un et à un seul indice de peptide i, et inversement l'ensemble des indices de fragments attachés aux indices de peptides forme une partition de l'ensemble $\{1 : L\}$.

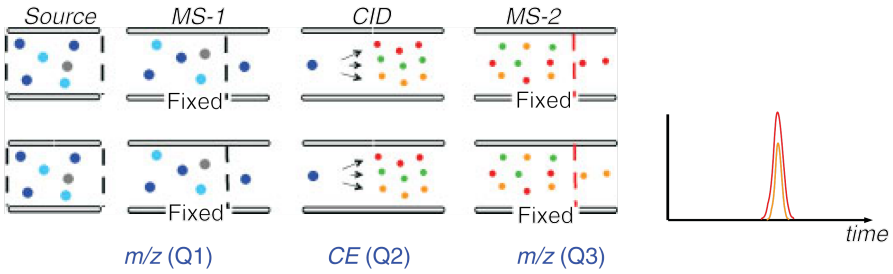
Pour résumer et écrire le modèle final d'une sortie SRM, nous faisons l'équivalence entre ion précurseur et peptide. Le signal modèle du fragment l enregistré par le détecteur qui est associé au peptide i a la forme

$$\boldsymbol{m}_l(t) = \beta_l \mathscr{C}_i(t; \tau_i, \lambda_i)\kappa_i. \tag{3.13}$$

Fig. 3.10 Principe du mode *Selected Reaction Monitoring* sur un triple quadrupole. L'ion précurseur est sélectionné par le premier quadrupole (Q1) servant de filtre de masse. Il passe ensuite dans la chambre de collision (Q2) où l'ion est fragmenté. Ensuite, un fragment est sélectionné par le deuxième filtre de masse (Q3). (b) montre la possibilité d'un monitoring de fragments multiples.



(a) Figure issue de [68]



(b) Figure issue de [69]

Ce signal est perturbé par un bruit blanc gaussien $\varepsilon_l(t;\gamma_l)$ d'inverse variance γ_l. On remarquera que le paramètre du bruit n'est pas global mais attaché à une trace. Entre traces, les niveaux de bruits peuvent en effet varier.

L'observation que l'on obtient en sortie est une version discrétisée du signal, et on a pour le fragment l la représentation vectorielle

$$\boldsymbol{y}_l = \beta_l \mathscr{C}_i \kappa_i + \boldsymbol{b}_l. \tag{3.14}$$

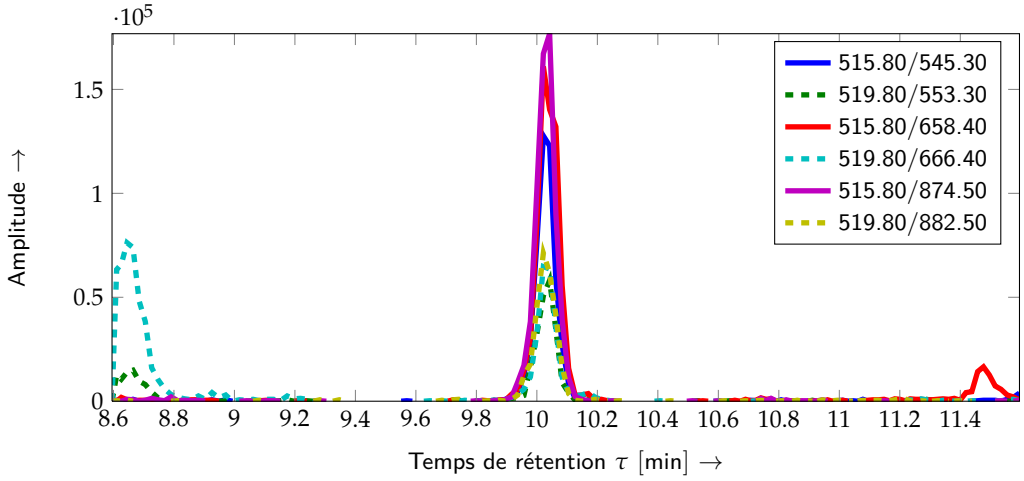
Remarque ⟨3.7⟩ *En général, les traces discrétisées d'un même peptide sont de taille identique puisqu'il s'agit du même contenu peptidique que l'on veut observer. A contrario, les traces de deux peptides différents ne le sont pas.*

Modèle d'une sortie LC-SRM avec marquage AQUA

Dans le cas des données SRM traitées dans cette thèse, le standard interne est le standard AQUA. Par ce moyen, tous les gains jusqu'à l'étage des peptides peuvent être surveillés facilement puisque les molécules natives et marquées ont les mêmes propriétés physico-chimiques. Cependant, lors de la présentation des données et des expérimentations, il nous est apparu que les traces pour les fragments d'un peptide n'ont pas toujours le même rapport entre amplitudes du natif et du marqué, même si ce phénomène est relativement rare. Le lecteur peut l'observer dans la Fig. 3.11 où le rapport entre les transitions 515.80/545.30 et 519.80/553.30 (bleu plein et vert pointillé respectivement) n'est pas le même que pour les autres couples « natif/marqué ». Pour modéliser cela, nous ajoutons un facteur multiplicatif d'ajustement $\phi_l^* \in \mathbb{R}_+$ dans l'expression de la trace marquée.

Le modèle de sortie d'un signal chromato-spectrométrique type SRM est donc donné

Fig. 3.11 Visualisation d'un spectre SRM d'un peptide GVSEIVQNGK natif (traits pleins), issu de la protéine L-FABP, et de sa variante marquée (traits pointillés). On voit bien que les traces ont la forme et la position des pics en commun. Aux bords de l'intervalle de temps de rétention apparaissent des pics parasites.



par le jeu d'équations suivant :

$$\boldsymbol{\kappa} = \chi \mathbf{D} \boldsymbol{\psi} \boldsymbol{x} + \varepsilon_{\boldsymbol{\kappa}}, \tag{3.15a}$$

$$\boldsymbol{y}_l = \beta_l \kappa_i \mathscr{C}(\tau_i, \lambda_i) + \boldsymbol{b}_l, \quad l = 1, \ldots, L, \tag{3.15b}$$

$$\boldsymbol{y}_l^* = \phi_l^* \beta_l \kappa_i^* \mathscr{C}(\tau_i, \lambda_i) + \boldsymbol{b}_l^*, \quad l = 1, \ldots, L. \tag{3.15c}$$

La sortie d'un signal type SRM est représenté dans la Fig. 3.11.

3.5 Modèle hiérarchique de sortie

Même si les instruments que l'on peut utiliser en protéomique sont de physique différente et ont donc des paramètres différents, on retrouve une structure hiérarchique commune. En effet, les modèles décrits précédemment – que ce soit pour une acquisition Full-MS ou SRM – sont des réalisations d'une cascade de processus qui peuvent s'écrire schématiquement dans un *arbre hiérarchique*, voir Fig. 3.12. Ceci nous donnera une certaine souplesse dans les développements, dans l'écriture, mais aussi dans la transposition à d'autres problèmes.

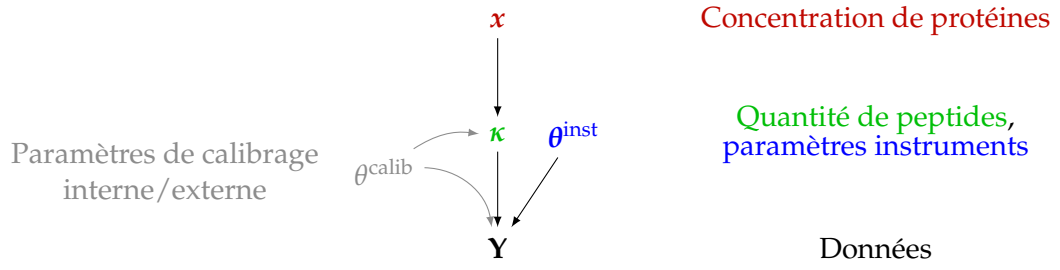
Par rapport aux études menées ici, nous pouvons identifier trois étages pour les paramètres considérés que nous allons exposer ci-dessous :

Premier étage : protéines. La variable de départ et d'intérêt dans la protéomique quantitative est la concentration de protéines $\boldsymbol{x}$. On verra plus loin dans ce document que la modélisation hiérarchique permettra de rajouter un étage en amont, correspondant au statut clinique et aux hyperparamètres de la distribution de la concentration.

Deuxième étage : peptides. Cet étage est séparé en deux branches :

1. La quantité de peptides qui est paramétrée par la concentration de protéines de l'étage précédent. C'est entre ces deux variables que l'on trouvera une dépendance hiérarchique.

Fig. 3.12 Graphe acyclique orienté générique du modèle d'acquisition



2. Les paramètres instruments qui dépendent de l'appareil choisi. Tel que nous avons conçu le modèle, il n'y a pas de dépendance *a priori* entre concentration d'une protéine et paramètres instruments.

Troisième étage : observations. On associe à cet étage les observations qui dépendent du deuxième étage entier : il s'agit d'une représentation des peptides acquis, éventuellement en passant par les fragments de peptides, à l'aide de l'instrument choisi.

Ce modèle générique peut ensuite être détaillé pour les modes Full-MS + PSAQ et SRM + AQUA en identifiant les paramètres instruments comme dans la Fig. 3.13 : les positions des pics chromatographiques, un paramètre d'inverse variance du bruit de mesure, les gains du système pour le premier, les positions et largeurs des pics chromatographiques, les paramètres d'inverse variance de bruit de mesure et les gains des traces. Chaque modèle comporte également le calibrage : au niveau des protéines pour les acquisitions Full-MS grâce au standard PSAQ ; au niveau des peptides pour les acquisitions SRM grâce au standard AQUA avec un précalibrage externe pour le passage « protéines/peptides », appelé *étalonnage par les échantillons de Contrôle de Qualité* (ou étalonnage CQ).

Les échantillons de Contrôle de Qualité sont généralement préparés dans les mêmes conditions et en même temps que les échantillons à analyser. Leur concentration protéique étant connue, le calibrage des gains – qui sont les seules inconnues de l'analyse – est ainsi rendu possible.

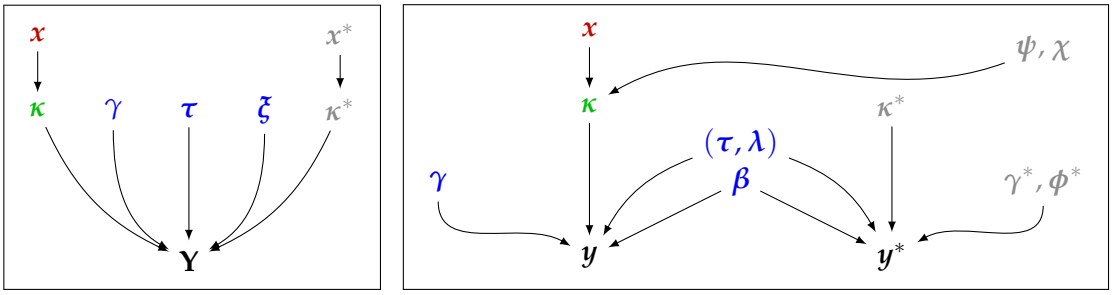
Le calibrage externe par CQ du couple de gain de préparation et rendement de digestion peut être contesté. En effet, les gains doivent être supposés stables d'une acquisition à l'autre pour permettre une réutilisation dans un autre échantillon du même jeu de préparation. Cette stabilité n'est probablement pas toujours satisfaite ; des moyens pour la surveiller sont en phase de réflexion par les équipes interagissant dans BHI-PRO.

Le contre-argument majeur pour notre application de l'étalonnage CQ est la dépendance d'une analyse extérieure à l'échantillon à traiter pour estimer la paramètre d'intérêt qui est la concentration protéique. L'**inversion** de la chaîne d'analyse se fait, comme on le verra, en ne considérant que les signaux de l'échantillon même. Les informations sont agrégées pour fournir une valeur [2] au plus haut niveau possible : celui du peptide. La valeur au niveau de la protéine s'obtient ensuite en appliquant une agrégation des valeurs peptidiques avec des « poids » appris sur les signaux des échantillons CQ, rapportant ainsi une variabilité du gain des CQ dans l'estimation de la concentration.

Néanmoins, nous utilisons cet étalonnage – avec précaution – à ce stade du développement.

2. Nous préférons le terme « valeur » à « estimation » car l'algorithme que nous développons dans la suite est un algorithme itératif, proposant des valeurs aléatoires pour chacun des paramètres selon une distribution de probabilité. L'agrégation des informations peptidiques se fait à chaque itération, avant de fournir une estimation.

Fig. 3.13 Graphes acycliques orientés du modèle d'acquisition suivant le mode de spéctrométrie de masse utilisé, reprenant le code couleur de Fig. 3.12, un paramètre gris correspondant à un paramètre de calibrage.



(a) Full-MS + PSAQ (b) SRM + AQUA

3.6 Description des données disponibles

La présente section décrit les données simulées et cliniques qui sont à notre disposition pour l'évaluation de nos développements, présentés dans la suite du document.

Suite à un effectif trop réduit de données réelles en mode Full-MS, nous nous contentons de simulations, l'intérêt d'une approche bayésienne étant démontrée dans [1] et [118]. Pour les données LC-SRM, nous disposons ensuite de trois jeux de données, dont deux de données réelles.

3.6.1 Full-MS : données simulées

Selon le modèle LC-Full-MS mis en place, nous simulons la mesure $\mathbf{Y}$ de la protéine *Entérotoxine A* du Staphylocoque doré (SEA) dont les propriétés et conditions d'acquisition que nous imitons ont été énumérées dans [1] et [118]. Ces acquisitions ont été réalisées sur le projet commun CAPSI avec le laboratoire EDyP (iRTSV, CEA Grenoble) [70]. Comme il s'agit d'une seule protéine, nous ne travaillerons plus avec les vecteurs et les matrices, mais avec des scalaires. Pour ce qui concerne les caractéristiques stables et non aléatoires dans notre cadre, voir Tab. 3.1.

En ce qui concerne les paramètres aléatoires, ils ont été simulés de la manière suivante. La concentration du biomarqueur x_n suit une loi normale de moyenne m et de précision γ. Ne suivant qu'un peptide, sa quantité κ_n suit une loi normale de moyenne x_n et de précision γ_κ. Ce peptide subit un gain d'ionisation ξ_n qui suit une loi normale de moyenne 10 et de précision 10. Le temps de rétention τ_n est choisi uniformément dans l'intervalle $[\tau^{\mathrm{m}}, \tau^{\mathrm{M}}] = [1628.5, 1668.5]$ s. Finalement, le niveau de bruit γ_n est tiré uniformément dans l'intervalle $[0.01, 1]$. Ainsi, nous pouvons simuler un grand effectif de données, certaines très perturbées par le bruit de mesure ou par d'autres processus aléatoires, ce qui permettra d'évaluer efficacement les méthodes pour les données LC-Full-MS.

3.6.2 SRM : données simulées

Pour valider la quantification de protéines à partir de données LC-SRM, nous mettons en place des simulations de données SRM selon le modèle présenté auparavant. Cela nous permet de confronter les valeurs estimées à des valeurs vraies que nous choisissons et maîtrisons. En s'inspirant des données réelles décrites ci-dessous, nous adaptons les paramètres de la génération de données pour obtenir un grand jeu d'échantillons

Tab. 3.1 Caractéristiques de la protéine SEA nécessaires pour la simulation de données LC-Full-MS.

- protéine
  - une protéine native, $P = 1$
    - $x_n \sim \mathcal{N}(m, \gamma)$ où m et γ varient selon le problème considéré
  - une protéine marquée (standard PSAQ)
- peptide
  - un peptide, $I = 1$
    - séquence : NVTVQELDLQAR
  - précision de digestion $\mathbf{\Gamma}_{\boldsymbol{\kappa}} = 1$
  - masses
    - $m = 1385.54\,\text{Da}$, $m^* = m + 9\,\text{Da}$
  - charges $J = 2$
    - $\pi_{1,1} = 0$
    - $\pi_{1,2} = 1$
  - neutrons $K = 0$
    - $\pi_{1,1,0} = 0$
    - $\pi_{1,2,0} = 1$
  - temps de rétention
    - $\tau \in [1628.5, 1668.5]$ s
  - gain d'ionisation
    - $\xi \sim \mathcal{N}(10, 10)$
- observations
  - largeur des pics
    - spectrométriques : $\lambda_{\mathscr{S}} = 0.0455^{-1}$
    - chromatographiques : $\lambda_{\mathscr{C}} = 14.08^{-1}$
  - fréquence d'échantillonnage
    - spectrométrique : $f^{\mathrm{e}}_{\mathscr{S}} = 1/M^{\mathrm{e}}_{\mathscr{S}} = 1/0.0425$
    - chromatographique : $f^{\mathrm{e}}_{\mathscr{C}} = 1/T^{\mathrm{e}}_{\mathscr{C}} = 1/3.339$
  - bruit de mesure
    - bruit blanc gaussien de moyenne nulle et de précision γ
    - $\gamma \sim \mathcal{U}([0.01, 1])$

biologiques pour l'évaluation et la validation des développements entrepris dans ce document.

3.6.3 SRM : données synthétiques

Le jeu de données utilisé pour la validation de la quantification à partir de données LC-SRM a été acquis par notre partenaire bioMérieux en décembre 2009 sur un spectromètre de masse AB Sciex QT5500 Triple Quadrupole. Il s'agit de données dites de « linéarité ». Ce jeu de données comporte une gamme de six échantillons de Contrôle de Qualité, notés CQ0, CQ1, ..., CQ5. Les échantillons biologiques sont obtenus en diluant un échantillon père, CQ5, à l'intérieur duquel se trouvent 60 protéines cibles à concentration connue. À chaque protéine sont associés 1 à 3 peptides, et à chaque peptide 2 à 4 transitions.

La dilution a lieu en puissances de 2 : le CQ5 a un rapport de dilution de $r_{\text{CQ5}} = 1 = 2^0$, le CQ4 de $r_{\text{CQ4}} = 2^{-1}$, ..., le CQ1 de $r_{\text{CQ1}} = 2^{-4}$. Quant au CQ0, il s'agit d'un échantillon « blanc » sans ajout des protéines cibles auquel on attribue le rapport de dilution $r_{\text{CQ0}} = 0$. Nous pouvons alors travailler avec des concentrations relatives par rapport à ces rapports de dilution.

Par CQ, nous disposons des données de trois injections, *i.e.* une par date de digestion (les 03, 10 et 11 décembre 2009). Suite à une analyse préalable des données, il s'avère que les données du 11 décembre semblent corrompues : la plupart des signaux sont vides au bruit de mesure près. Nous excluons donc ce jour de digestion dans nos évaluations, ce qui nous permet de « conclure » à partir d'un jeu de données à deux injections par CQ, *i.e.* 12 acquisitions. Strictement parlant, ce nombre est trop peu élevé pour en déduire des résultats statistiquement valables, mais nous pouvons prendre en compte les tendances mises en évidence.

3.6.4 SRM : données du cancer colorectal

Entre janvier et mars 2010, bioMérieux a collecté un jeu de données réelles d'une cohorte de 237 individus anonymisés, acquis sur un spectromètre de masse AB Sciex QT5500 Triple Quadrupole. La cohorte peut être séparée en deux classes :

1. **Contrôle**. Cette classe de 122 individus possède deux sous-classes.

  - EFS. Il s'agit de 91 données acquises à partir d'échantillons biologiques de donneurs de sang de l'Établissement Français du Sang de Lyon, ayant autorisé l'utilisation de leur don à des fins de recherches scientifiques.
  - Coloscopie Négative. Ces 31 échantillons proviennent d'acquisitions d'individus non atteint du cancer colorectal, selon l'analyse de coloscopie.
2. **Cas**. Dans cette classe de 115 patients du cancer colorectal, nous distinguons quatre stades du développement du cancer, *i.e.* sous-classes.
  - Stade 1, 27 échantillons.
  - Stade 2, 32 échantillons.
  - Stade 3, 29 échantillons.
  - Stade 4, 27 échantillons.

Finalement, la cohorte analysée comporte (sans réplication de mesures) les données de 203 patients distincts, dont 11 sans classe identifiée.

Afin d'exploiter les données dans l'environnement de travail, MATLAB, les signaux d'une acquisition sont d'abord extraits d'un fichier de format propriétaire (extension `.wiff`) qui en regroupe plusieurs. Ils sont enregistrés dans un fichier de format libre [71] (extension `.mzml`) lisible dans MATLAB à l'aide d'une routine de lecture et conversion. Ensuite, afin de permettre une utilisation conforme à nos développements, les caractéristiques de l'acquisition sont recueillies et enregistrées avec les signaux dans un format natif MATLAB (extension `.mat`). Parmi les caractéristiques, nous comptons notamment le nom du fichier `.wiff` propagé dès le début, les appellations des protéines, les séquences des peptides, les masses des transitions, le nombre des molécules, les associations protéines/peptides/fragments, *et cetera*.

Dans chaque acquisition convertie pour une fraction donnée, 40 protéines ont finalement été ciblées, donnant environ 180 transitions (1 à 4 peptides par protéine, 2 à 4 transitions par peptide). Contrairement au jeu de données linéarité, pour le jeu du cancer colorectal aucune valeur nominale « vraie » de la concentration protéique n'est connue. Cependant, il existe des mesures immuno-enzymatiques (*Enzyme Linked Immunosorbent Assay*), appelées ci-après ELISA, pour la protéine *Liver-type fatty-acid-binding protein*, abrégé LFABP. Il s'agit également d'une *estimation* de la concentration protéique à partir d'une mesure de la fluorescence de la réaction de la protéine, réputée fiable et performante. Son inconvénient repose clairement dans son coût de développement : actuellement, bioMérieux a la possibilité de développer six méthodes ELISA par an seulement, lié au temps de développement des anticorps de reconnaissance.

Nous comparons dans le Ch. 4 nos estimations issues de l'algorithme d'Inversion-Quantification avec les résultats de l'ELISA pour la protéine LFABP, tout en gardant à l'esprit qu'il s'agit d'une comparaison de deux estimations. Le livre [72] propose de corriger cette incertitude ce qui nécessite la connaissance de la variance de chaque jeu d'estimations. À l'instant où le présent manuscrit est rédigé, nous ne disposons pas de mesure de la variabilité des tests ELISA. Nous sommes donc contraints de les considérer comme référence exacte.

La cohorte du cancer colorectal est ensuite utilisée pour évaluer la performance de la classification, exposée dans le Ch. 5, en comparant l'estimation de la classe avec l'étiquette donnée au préalable.

3.7 Modélisation probabiliste

L'écriture hiérarchique du modèle « physique » direct que nous avons vu dans la Sect. 3.5 nous permet de transmettre les notions de dépendances entre étages à une modélisation dite « probabiliste ». Cette dernière donne une vision aléatoire du modèle direct

et prépare l'inversion dans un cadre bayésien.

Tous les processus que nous avons cités ont une certaine incertitude suite aux variabilités et aux fluctuations. Nous avons déjà introduit des termes d'erreurs dans la modélisation physique pour simuler ces incertitudes, et nous avons également constaté que certains paramètres sont instables d'une expérience à l'autre comme les gains et paramètres de bruit. Nous allons intégrer ces connaissances pour exploiter le signal observé dans ce qui suit.

Pour bien expliquer les processus aléatoires qui surviennent dans les modèles décrits ci-dessus, il convient de choisir les distributions adéquates. Le durée du traitement numérique dépend également du choix des lois *a priori*. Dans la Sect. 2.9.1, nous avons déjà motivé le choix de lois conjuguées quand ceci est possible.

En effet, certains paramètres sont connus comme par exemple la matrice de digestion $\mathbf{D}$ qui exprime le nombre de copies du peptide i dans la protéine p ou le profil spectrométrique puisque la distribution de la masse de l'ion est connue. Ce sont les paramètres dits stables. Leurs valeurs sont connues par expertise ou par calibrage. L'estimation n'est pas nécessaire pour eux.

Dans la suite, nous proposons des lois *a priori* pour les paramètres instables. Ce sont les paramètres qui ne sont pas connus avant l'expérience, avant le passage de l'échantillon biologique dans la chaîne analytique. Il s'agit principalement de la concentration des protéines cibles, la quantité de peptides, les gains divers traçables grâce à l'étalonnage, les paramètres des pics chromatographiques, et les paramètres d'inverse variance de bruit.

3.7.1 Paramètres communs

Concentration de protéines

La concentration des protéines étudiées est distribuée sous une loi normale de moyenne $\boldsymbol{m}_x$ et de matrice de précision $\boldsymbol{\Gamma}_x$. Il y a plusieurs raisons pour ce choix. Premièrement, il apparaît que la distribution normale pour ce paramètre est proche de ce qui a été constaté avec les tests ELISA. Deuxièmement, il s'agit d'une loi « flexible », c'est-à-dire que nous pouvons choisir d'ajouter de l'information en indiquant bien la moyenne et la précision, ou bien de choisir une précision faible pour ajouter le moins d'information possible, la distribution devenant plus large et favorisant moins les valeurs autour de la moyenne. Dans le cas extrême où la précision tend vers zéro, on approche une loi uniforme sur tout le support de la distribution qui est $\mathbb{R}$ [3]. Troisièmement, la distribution normale est plus facile à intégrer dans les développements algorithmiques du fait des possibilités de conjugaison de lois qu'elle propose et de la linéarité, contrairement à d'autres lois comme la loi uniforme ou la loi laplacienne.

Finalement, on a choisi comme modèle de distribution :

$$\boldsymbol{x} \sim \mathrm{p}_x(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_x, \boldsymbol{\Gamma}_x). \tag{3.16}$$

Quantité de peptides

Le passage de protéines en peptides a été modélisé à l'aide de la matrice de digestion $\mathbf{D}$, perturbé par un bruit, voir Éqn. (3.3). Compte tenu des connaissances sur le bruit,

3. Strictement parlant, nous devrions choisir une loi qui a un support positif. On pourrait par exemple choisir une loi normale tronquée sur $[0, +\infty[$. L'implantation de la loi normale (non tronquée) est cependant plus facile à mettre en œuvre ce qui motive le choix de cette dernière. Il reste cependant à surveiller la positivité des variables résultantes. Cette même remarque tient également pour tous les paramètres qui suivent puisqu'il s'agit de paramètres physiques positifs ou nuls.

nous modélisons le bruit comme un processus blanc gaussien de moyenne nulle et de précision $\boldsymbol{\Gamma}_{\boldsymbol{\kappa}}$. Ainsi, le processus de digestion peut être décrit par

$$\boldsymbol{\kappa} \sim \mathrm{p}_{\kappa}(\boldsymbol{\kappa} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{\kappa}; \Delta \boldsymbol{x}, \boldsymbol{\Gamma}_{\boldsymbol{\kappa}}). \tag{3.17}$$

Dans le cas PSAQ, la matrice Δ correspond pleinement à la matrice de digestion $\mathbf{D}$, les gains étant pris en compte par la variable gain de système ; pour l'AQUA, Δ vaut $\boldsymbol{\chi}\mathbf{D}\boldsymbol{\psi}$.

Notons que pour décrire une digestion peu bruitée, il suffit de choisir une précision très grande. Le cas limite est atteint lorsque la précision est infinie : le lien modélisé du processus devient déterministe.

3.7.2 Paramètres instruments du couplage LC-Full-MS

Position des pics chromatographiques

Les positions des pics chromatographiques peuvent varier d'une expérience à l'autre. Il est tout de même possible de déterminer une plage de valeurs possibles, par les connaissances de l'expérimentateur, par une base de donnée, par expériences d'étalonnage, ou par un calcul qui fait intervenir plusieurs propriétés du peptide concerné [73]. Cependant, à l'intérieur de cet intervalle, nous ne pouvons formuler de préférence pour une valeur ou pour une autre. En admettant que cet intervalle $[\tau^{\mathrm{m}}, \tau^{\mathrm{M}}]$ existe, nous proposons d'utiliser une loi uniforme qui traduit le mieux la variation du temps de rétention :

$$\boldsymbol{\tau} \sim \mathrm{p}(\boldsymbol{\tau}) = \mathcal{U}(\boldsymbol{\tau}; \boldsymbol{\tau}^{\mathrm{m}}, \boldsymbol{\tau}^{\mathrm{M}}). \tag{3.18}$$

Gains de système

La loi pour les gains est obtenue de façon similaire à celle de la concentration des protéines. On leur attribue une loi normale pour leurs bonnes propriétés statistiques :

$$\boldsymbol{\xi} \sim \mathrm{p}(\boldsymbol{\xi}) = \mathcal{N}(\boldsymbol{\xi}; \boldsymbol{m}_{\xi}, \boldsymbol{\Gamma}_{\xi}). \tag{3.19}$$

Si on décide d'injecter peu d'information quant à la valeur du gain du peptide i donné, on fait tendre sa précision $[\boldsymbol{\Gamma}_{\xi}]_{ii}$ vers une valeur faible $\varepsilon > 0$. Ceci n'empêche nullement les autres composantes à contenir de l'information par rapport à la variance des gains.

Inverse variance du bruit de mesure

Définissons maintenant la loi de probabilité pour l'inverse variance du bruit. Le bruit de mesure étant modélisé gaussien et donc la vraisemblance totale étant de densité normale, nous choisissons une loi gamma pour ce paramètre de bruit. Ce choix permettra la conjugaison de la loi *a priori* par la vraisemblance. Dans la littérature, il existe deux définitions équivalentes pour cette loi (cf. [2, Sect. A.2]). Nous utiliserons celle qui fait intervenir un paramètre de forme $\alpha > 0$ et un paramètre d'échelle $\beta > 0$ (*i.e.* la deuxième convention de la citation précédente correspondant à celle que donne [38, App. 1]) :

$$\gamma \sim \mathrm{p}(\gamma) = \mathcal{G}(\gamma; \alpha, \beta) = \gamma^{\alpha-1} \frac{\exp\left(-\frac{\gamma}{\beta}\right)}{\beta^{\alpha}\Gamma(\alpha)} \cdot \mathbb{1}_{[0,\infty[}(\gamma). \tag{3.20}$$

Bruit et vraisemblance totale

Nous avons modélisé le bruit de mesure comme étant blanc gaussien, centré, de moyenne nulle et d'inverse variance γ pour chaque point de mesure :

$$\mathbf{B}[m,n] \sim \mathrm{p}(\mathbf{B}[m,n] \mid \gamma) = \mathcal{N}(\mathbf{B}[m,n]; 0, \gamma). \tag{3.21}$$

Les réalisations de bruit sur les N points de mesures étant indépendantes, la densité de probabilité pour la matrice du bruit peut s'écrire de la manière suivante :

$$\begin{aligned}
\mathbf{B} \sim \mathrm{p}(\mathbf{B} \mid \gamma) &= \prod_{m=1}^{N_{\mathscr{S}}} \prod_{n=1}^{N_{\mathscr{C}}} \mathrm{p}(\mathbf{B}[m,n] \mid \gamma) \\
&= \prod_{m=1}^{N_{\mathscr{S}}} \prod_{n=1}^{N_{\mathscr{C}}} \mathcal{N}(\mathbf{B}[m,n]; 0, \gamma) \\
&= (2\pi)^{-N/2} \gamma^{N/2} \exp\left(-\frac{\gamma}{2} \sum_{m=1}^{N_{\mathscr{S}}} \sum_{n=1}^{N_{\mathscr{C}}} \mathbf{B}[m,n]^2\right) \\
&= (2\pi)^{-N/2} \gamma^{N/2} \exp\left(-\frac{\gamma}{2} \|\mathbf{B}\|_{\mathrm{F}}^2\right)
\end{aligned} \tag{3.22}$$

où $N = N_{\mathscr{S}} \cdot N_{\mathscr{C}}$ est le nombre total de points de mesure et $\|\cdot\|_{\mathrm{F}}$ désigne la norme de Frobenius[xviii] d'une matrice. Cette dernière correspond à la norme euclidienne dans un espace de dimension $N_{\mathscr{S}} \cdot N_{\mathscr{C}}$, la matrice étant vectorisée. Pour une matrice $\mathbf{A} \in \mathbb{C}^{N_{\mathscr{S}} \times N_{\mathscr{C}}}$, elle est définit par $\|\mathbf{A}\|_{\mathrm{F}} = \mathrm{tr}(\mathbf{A}\mathbf{A}^{\mathrm{H}})^{1/2} = \left(\sum_{s=1}^{N_{\mathscr{S}}} \sum_{c=1}^{N_{\mathscr{C}}} |a_{sc}|^2\right)^{1/2}$.

Par une reparaméterisation en tenant compte du fait que $\mathbf{Y} = \mathbf{M} + \mathbf{M}^* + \mathbf{B}$, nous écrivons la loi du résidu de la même manière :

$$\begin{aligned}
\mathbf{Y} - \mathbf{M} - \mathbf{M}^* = \mathbf{B} \sim \mathrm{p}(\mathbf{B} \mid \gamma) &= (2\pi)^{-N/2} \gamma^{N/2} \exp\left(-\frac{\gamma}{2} \|\mathbf{B}\|_{\mathrm{F}}^2\right) \\
&= (2\pi)^{-N/2} \gamma^{N/2} \exp\left(-\frac{\gamma}{2} \|\mathbf{Y} - \mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2\right).
\end{aligned} \tag{3.23}$$

La loi du résidu est identique à la loi de probabilité des données connaissant les sorties modèle qui sont déterminées par la quantité des peptides $\boldsymbol{\kappa}$ et les paramètres instruments $\boldsymbol{\theta}^{\mathrm{inst}}$. Ce terme est donc un terme d'attache ou de fidélité aux données. Ainsi, la densité

$$\mathrm{p}(\mathbf{Y} \mid \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}}) = (2\pi)^{-N/2} \gamma^{N/2} \exp\left(-\frac{\gamma}{2} \|\mathbf{Y} - \mathbf{M} - \mathbf{M}^*\|_{\mathrm{F}}^2\right) = \mathcal{N}(\mathbf{Y}; \mathbf{M} + \mathbf{M}^*, \gamma) \tag{3.24}$$

est la fonction de vraisemblance totale, c'est-à-dire la vraisemblance des paramètres attachées aux données.

3.7.3 Paramètres instruments du couplage LC-SRM

Position et largeur des pics chromatographiques

Généralement, la position d'un pic chromatographique est connue dans une certaine plage de temps, sans avoir d'indication plus exacte pour un endroit particulier. Cette connaissance est exploitée notamment dans les modes d'acquisition *Scheduled SRM* où les transitions ne sont obtenues et enregistrées qu'à des temps chromatographiques définis. Il convient alors de proposer une distribution uniforme dans un intervalle pour la position du pic chromatographique pour chaque peptide. En admettant une écriture multidimensionnelle de la distribution avec un intervalle multidimensionnel $\left[\boldsymbol{\tau}^{\mathrm{m}}, \boldsymbol{\tau}^{\mathrm{M}}\right]$, on a

$$\boldsymbol{\tau} \sim \mathrm{p}(\boldsymbol{\tau}) = \mathcal{U}(\boldsymbol{\tau}; \boldsymbol{\tau}^{\mathrm{m}}, \boldsymbol{\tau}^{\mathrm{M}}). \tag{3.25}$$

Suite aux expériences et à l'expertise que nous avons acquise lors des analyses de données, nous pouvons extraire une plage de valeurs possibles pour la largeur du pic chromatographique. Ainsi, la largeur de chaque pic est distribuée uniformément dans un intervalle multidimensionnel $\left[\lambda^{\mathrm{m}}, \lambda^{\mathrm{M}}\right]^{I}$; nous pouvons écrire

$$\boldsymbol{\lambda} \sim \mathrm{p}(\boldsymbol{\lambda}) = \mathcal{U}\left(\boldsymbol{\lambda}; \left[\lambda^{\mathrm{m}}, \lambda^{\mathrm{M}}\right]^{I}\right). \tag{3.26}$$

L'intervalle typique des largeurs, visant l'exclusion de profils trop pointus ou trop étendus, est $\left[\lambda^{\mathrm{m}}, \lambda^{\mathrm{M}}\right] = [20,\ 100]\,\mathrm{min}^{-2}$, ce qui revient à proposer une déviation du pic chromatographique entre 0.1 et 0.22 min.

Gains de fragmentation

Les valeurs pour les gains de fragmentation $\boldsymbol{\beta}$ ne sont pas connues *a priori*. Pour retranscrire cela, nous pouvons utiliser une loi uniforme sur un intervalle très large. Cependant, au vu de la conjugaison par la vraisemblance qui est normale comme on le verra dans la suite, nous préférons une distribution normale avec la moyenne $\boldsymbol{m}_{\beta}$ et la précision $\boldsymbol{\Gamma}_{\beta}$:

$$\boldsymbol{\beta} \sim \mathrm{p}(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{m}_{\beta}, \boldsymbol{\Gamma}_{\beta}). \tag{3.27}$$

Le manque de connaissance *a priori* pourra être décrit par une précision faible, créant une loi vaguement informative.

Gains de transitions

Pour une transition l donnée, suite au partage des paramètres instruments, le rapport entre les amplitudes des traces du peptide natif et du marqué est censé être égal au rapport entre les quantités peptidiques du natif du marqué. Autrement dit, le paramètre ϕ_l^* qui peut être interprété comme rapport de ces rapports doit valoir 1, à l'incertitude près. Ce fait peut être traduit par au moins deux distributions :

(1) Une distribution gamma de paramètres de forme α et d'échelle $\beta = \alpha^{-1}$, la moyenne de cette distribution étant $\alpha\beta = 1$, son mode $(\alpha - 1)\beta = 1 - \alpha^{-1}$ et son support étant $\mathbb{R}_+$. Plus α est grand, plus le mode d'approche de 1, et plus la distribution est concentrée autour du mode.

(2) Une distribution normale de moyenne $m_{\phi^*} = 1$ et de précision γ_{ϕ^*}. Plus la précision est grande, plus la distribution est concentrée autour de 1.

La distribution normale étant conjuguée par la vraisemblance associée, c'est cette dernière que nous choisissons malgré le support $\mathbb{R}$. Elle permettra d'utiliser un générateur d'échantillons aléatoires « standards » sans devoir passer par une étape coûteuse de Metropolis-Hastings qui aurait été nécessaire en choisissant la distribution gamma.

Inverse variance du bruit des mesures

D'après la modélisation ci-dessus, le bruit de mesure de chaque trace est blanc, gaussien. Pour des questions de conjugaison, ceci motive le choix de la loi gamma de forme α_l et d'échelle β_l comme distribution du paramètre de bruit γ_l pour la trace du fragment l. Il en est de même pour les traces des peptides marqués. Alors, pour chaque trace $l = 1, \ldots, L$,

$$\gamma_l \sim \mathrm{p}(\gamma_l) = \mathcal{G}(\gamma_l; \alpha_l, \beta_l), \tag{3.28}$$

$$\gamma_l^* \sim \mathrm{p}(\gamma_l^*) = \mathcal{G}(\gamma_l^*; \alpha_l^*, \beta_l^*). \tag{3.29}$$

Bruit et vraisemblance totale

Dans la section de la modélisation directe d'une sortie SRM, nous avons décidé de modéliser le bruit de mesure d'une trace par un bruit blanc gaussien, de moyenne nulle et d'inverse variance γ_l pour la trace d'une molecule native.

$$\boldsymbol{b}_l[n] \sim \mathrm{p}(\boldsymbol{b}_l[n] \mid \gamma_l) = \mathcal{N}(\boldsymbol{b}_l[n]; 0, \gamma_l). \tag{3.30}$$

Les points de mesures étant conditionnellement indépendants entre eux, la distribution du vecteur de bruit est le produit des distributions de chaque point et s'écrit

$$\begin{aligned}\mathrm{p}(\boldsymbol{b}_l \mid \gamma_l) &= \prod_{n=1}^{N_l} \mathrm{p}(\boldsymbol{b}_l[n] \mid \gamma_l) = \prod_{n=1}^{N_l} \mathcal{N}(\boldsymbol{b}_l[n]; 0, \gamma_l) \\ &= (2\pi)^{-N_l/2} \gamma_l^{N_l/2} \exp\left(-\frac{\gamma_l}{2} \boldsymbol{b}_l^{\mathrm{T}} \boldsymbol{b}_l\right)\end{aligned} \tag{3.31}$$

où N_l est le nombre de points de mesure de la trace l.

Par un changement de paramètres, nous obtenons facilement de la loi du bruit celle de l'observation connaissant la quantité peptidique κ_i attachée au fragment l et les paramètres instruments $\boldsymbol{\theta}_l^{\mathrm{inst}}$, compte tenu du fait que $\boldsymbol{y}_l = \beta_l \kappa_i \mathscr{C}(\tau_i, \lambda_i) + \boldsymbol{b}_l = \boldsymbol{m}_l + \boldsymbol{b}_l$ (voir Éqn. (3.15)),

$$\mathrm{p}(\boldsymbol{y}_l \mid \kappa_i, \boldsymbol{\theta}_l^{\mathrm{inst}}) = (2\pi)^{-N_l/2} \gamma_l^{N_l/2} \exp\left(-\frac{\gamma_l}{2} (\boldsymbol{y}_l - \boldsymbol{m}_l)^{\mathrm{T}} (\boldsymbol{y}_l - \boldsymbol{m}_l)\right). \tag{3.32}$$

Nous venons de déduire la distribution de la trace l, sachant les paramètres peptides et instruments intervenant dans la génération de cette dernière. Les acquisitions étant indépendantes, la distribution de toutes les traces natives est donnée par le produit des distributions des traces natives :

$$\mathrm{p}(\boldsymbol{y}_{1:L} \mid \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}}) = \prod_{l=1}^{L} \mathrm{p}(\boldsymbol{y}_l \mid \kappa_i, \boldsymbol{\theta}_l^{\mathrm{inst}}). \tag{3.33}$$

Le développement précédent s'applique également aux traces marquées, en utilisant la modélisation adéquate Éqn. (3.15). En introduisant la quantité peptidique du standard AQUA $\boldsymbol{\kappa}^*$ — supposée fixe et ajoutée dans le conditionnement à des fins d'exhaustivité —, on a

$$\mathrm{p}(\boldsymbol{y}_{1:L}^* \mid \boldsymbol{\kappa}^*, \boldsymbol{\theta}^{\mathrm{inst}}, \boldsymbol{\phi}^*) = \prod_{l=1}^{L} \mathrm{p}(\boldsymbol{y}_l^* \mid \kappa_i^*, \boldsymbol{\theta}_l^{\mathrm{inst}}, \phi_l^*). \tag{3.34}$$

Réunissons maintenant les distributions définies dans Éqn. (3.33) et Éqn. (3.34). Le produit de ces deux distributions est la distribution de l'ensemble des observations, sachant tous les paramètres de génération des données :

$$\mathrm{p}(\boldsymbol{y}_{1:L}, \boldsymbol{y}_{1:L}^* \mid \boldsymbol{\kappa}, \boldsymbol{\kappa}^*, \boldsymbol{\theta}^{\mathrm{inst}}, \boldsymbol{\phi}^*) = \mathrm{p}(\boldsymbol{y}_{1:L} \mid \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}}) \cdot \mathrm{p}(\boldsymbol{y}_{1:L}^* \mid \boldsymbol{\kappa}^*, \boldsymbol{\theta}^{\mathrm{inst}}, \boldsymbol{\phi}^*) \tag{3.35}$$

qui est, du point de vue des paramètres peptide $\boldsymbol{\kappa}$, instruments $\boldsymbol{\theta}^{\mathrm{inst}}$ et calibrants $\boldsymbol{\kappa}^*$ et $\boldsymbol{\phi}^*$, la fonction de vraisemblance totale. Elle mesure l'attache de l'ensemble des paramètres du deuxième étage hiérarchique (Sect. 3.5) par rapport aux observations disponibles.

3.8 Bilan

Dans ce chapitre, nous avons modélisé la chaîne analytique adaptée pour l'analyse d'échantillons biologiques. Pour cela, nous avons notamment présenté les étapes de préparation, la colonne de chromatographie et le spectromètre de masse avec les deux modes

de lecture considérés dans cette thèse. Nous avons en particulier mis en avant le caractère hiérarchique de cette chaîne. Ensuite, nous en avons déduit les modèles directs physique et probabiliste avec la proposition de lois *a priori* pour les paramètres. Ainsi, nous avons posé les bases pour les étapes d'inversion que nous étudierons dans le Ch. 4 pour l'Inversion-Quantification de la concentration et dans le Ch. 5 pour l'Inversion-Classification d'un échantillon biologique à statut clinique indéterminé.

4 Inversion-Quantification

Dans ce chapitre, nous présentons la quantification de protéines à l'aide de données chromato-spectrométriques. Pour ce faire, nous proposons l'utilisation du cadre des problèmes inverses et résolvons le problème posé à l'aide de méthodes statistiques bayésiennes.

Dans cette thèse, nous avons pu traiter les données de deux modes de spectrométrie de masse. Comme nous avons vu précédemment, les chaînes d'analyse se rejoignent à certains endroits, soit par l'utilisation des mêmes paramètres, soit par le possible regroupement de paramètres dans un niveau hiérarchique. Le regroupement permet ensuite de généraliser les calculs, mais aussi de simplifier et clarifier l'écriture des développements. C'est la raison pour laquelle nous raisonnons le plus possible avec des paramètres *hiérarchiques*, de *nuisance* et d'*intérêt* dont nous donnons les définitions ci-dessous dans le texte.

4.1 État de l'art

L'introduction de l'étalonnage a ouvert les portes de la protéomique quantitative « simple à réaliser » par rapport à des méthodes sans marquage (Sect. 3.1.4). [↓1] Si on relie l'intensité d'une trace à la concentration, il faut d'abord définir comment on veut en déterminer l'intensité. Nous résumons trois méthodes, deux utilisées classiquement en protéomique, la dernière faisant intervenir une modélisation probabiliste bayésienne.

4.1.1 Maximum du pic

La manière la plus simple pour estimer l'intensité d'un signal est de considérer son maximum. Dans le cas d'utilisation de mode Full-MS, nous obtenons un signal 2D. Pour préparer le calcul du maximum du pic, nous projetons une région 2D bien définie du chromato-spectrogramme sur l'axe chromatographique pour obtenir un *chromatogramme extrait* (appelé couramment XIC pour *Extracted Ion Current*). Cette région est définie comme suit : pour un temps chromatographique donnée, nous cumulons les valeurs sur l'axe spectrométrique à l'intérieur de l'intervalle $[m_1 - \tilde{\sigma}_-; m_1 + \tilde{\sigma}_+]$ où

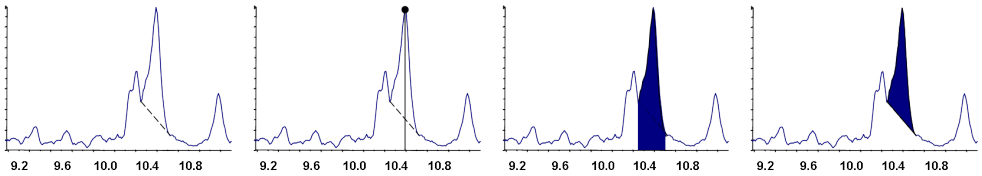
- m_1 est le rapport masse sur charge du premier pic du massif isotopique de la molécule d'intérêt,
- $\tilde{\sigma}_-$ et $\tilde{\sigma}_+$ sont respectivement l'étendue de l'intervalle vers la gauche et vers la droite pour couvrir le signal.

On peut par exemple choisir $\tilde{\sigma}_- = \tilde{\sigma}_+ = 2\lambda_{\mathscr{S}}^{-1}$ en fonction de la largeur du pic spectrométrique $\lambda_{\mathscr{S}}$ pour couvrir environ 95% du signal selon notre modélisation. Ensuite, on extrait un deuxième XIC pour la molécule marquée.

De chacun des deux signaux 1D (XIC ou signal SRM), nous extrayons le maximum qui est représentatif de la concentration de la molécule. Finalement, sous l'hypothèse que la forme des pics des signaux marquées et non marqués est la même et que le bruit est

1. Notons cependant qu'il existe des méthodes de quantification dites « label free » : elles se passent du marquage isotopique interne et calibrent par rapport à d'autres caractéristiques, comme des protéines stables (voir p.ex. [74–78]).

Fig. 4.1 Différentes méthodes d'intégration pour un signal donné.
(a) Représentation du signal considéré ; (b) Maximum du pic ; (c) Aire sous le pic négligeant la ligne de base ; (d) Aire sous le pic avec respect d'une ligne de base $L(S)$ linéaire entre les deux minima entourant le pic d'intérêt.



(a) (b) (c) (d)

négligeable, le rapport entre la concentration du marqué et le maximum du pic du signal XIC marqué est égal au rapport entre la concentration du natif et le maximum du pic du signal XIC natif. La seule inconnue dans cette équation étant la concentration du natif, il est facile de l'en déduire.

L'approche présentée n'utilise qu'un point de mesure pour déterminer l'intensité du pic. Or, moins on utilise de données, moins l'approche est robuste au bruit ce qui veut dire que la méthode du Maximum du pic est certes simple, mais peu robuste.

4.1.2 Aire sous le pic

L'extension naturelle de la méthode précédente est l'utilisation de l'aire sous le pic à partir du XIC, éventuellement après filtrage des données. Une fois l'intervalle d'intérêt sur l'axe chromatographique déterminé, l'aire calculée est représentative de la concentration de la molécule cible. Mais où est-ce que le pic commence, où est-ce qu'il s'arrête ? Une fois l'intervalle $[a, b]$ trouvé, comment intégrer la courbe (voir Fig. 4.1) ? Chez certains auteurs, la ligne de base est négligée, alors $M = \int_a^b S(t)\,\mathrm{d}t$ où S est le signal XIC ; chez d'autres, elle est importante et soustraite, alors $M = \int_a^b S(t) - L(S)(t)\,\mathrm{d}t$ où $L(S)$ est la ligne de base du signal. Quelle que soit la méthode d'intégration choisie, on procède comme pour le maximum du pic en comparant la mesure d'intensité du marqué à celle du natif.

Remarque ⟨4.1⟩ *Dans le contexte précédent, soit $\mu_\delta(t) = \delta_{t_{\max}}(t) = \delta_0(t - t_{\max})$ la mesure de Dirac translatée qui prend la masse 1 au point $t = t_{\max}$ et 0 sinon, et soit $t_{\max}$ la position du maximum du pic considéré. La quantification par le maximum du pic peut ainsi être interprétée comme intégration du signal sur l'intervalle $[a, b]$ par rapport à cette mesure de Dirac : $M = \int_a^b S(t)\mu_\delta(t) = \int_a^b S(t)\delta_0(t - t_{\max})\,\mathrm{d}t = (S * \delta_0)(t_{\max}) = S(t_{\max})$. Ce calcul fait apparaître la convolution du signal par la fonction de Dirac translatée.*

4.1.3 Quantification bayésienne par inversion

Alors que les deux méthodes présentées ci-dessus se passent d'une modélisation des données, l'inclusion des informations sur l'instrument par un modèle paramétrique a pour objectif d'améliorer l'estimation des intensités. La thèse de G. Strubel [1] propose d'estimer la concentration à l'aide d'une inversion bayésienne paramétrique des mesures. Après un travail sur la physique du tandem LC-MS, le manuscrit de ladite thèse propose une modélisation directe de l'instrument et l'inversion par utilisation des méthodes statistiques bayésiennes. Pour les calculs des densités de probabilité et des estimateurs, l'auteur choisit les méthodes MCMC en intégrant une structure de Gibbs hybride. La

performance et la robustesse de cette inversion sont démontrées dans la thèse citée, mais aussi dans les travaux [18, 79] ainsi que [118] qui y sont liés.

Remarque ⟨4.2⟩ *La thèse de G. Strubel préparée au CEA Leti a naturellement inspiré les travaux présentés dans ce document qui sont la suite logique de trois années fructueuses qui ont laissé un certain nombre de perspectives.*

4.1.4 Conclusion

Nous avons vu des méthodes courantes voire récente et originale pour l'estimation de la concentration de protéines. Nous pouvons cependant constater quelques défauts. Les premières méthodes n'exploitent aucunement les informations que nous avons à disposition quant à l'instrument. De plus, la première est très sensible au bruit de mesure et, le cas échéant, à l'estimation manuelle de la position et de l'amplitude des pics. La deuxième, quoique moins sensible au bruit, l'est par rapport à des pics contaminants qui perturbent la détermination de l'intervalle du pic qui est nécessaire pour le calcul de l'aire.

Quant à la troisième méthode, alors que l'auteur de [1] introduit la notion de hiérarchie dans l'acquisition des données (le sens *direct* du problème) en présentant la chaîne d'analyse comme succession de processus, il ne l'exploite ni dans la construction de son modèle probabiliste ni dans l'inversion. Le modèle probabiliste sous-jacent est non hiérarchique, *i.e.* par rapport à notre modélisation de base, le lien *protéines* $\rightarrow$ *peptides* est déterministe. L'ajout d'autres paramètres est bien évidemment possible dans le cadre de la référence citée, mais la gestion devient de plus en plus compliquée. Il semble naturel d'étendre ces travaux à des structures hiérarchiques qui peuvent aussi exprimer des liens déterministes [2], nous permettant aussi facilement de gérer le standard AQUA et l'étalonnage CQ dans le cas de l'utilisation du mode SRM.

Nous verrons dans la suite de ce document que l'ajout de variables dans un cadre hiérarchique est moins compliqué à réaliser. Un autre avantage est la simplification des calculs au vu des indépendances conditionnelles entre niveaux contrairement au cadre non hiérarchique. Enfin, utiliser un modèle hiérarchique veut aussi dire être plus proche de la réalité des processus ce qui permet de rendre plus robuste l'estimation par analyse bayésienne.

4.2 Modèle direct et loi jointe

Dans cette section, en utilisant le modèle hiérarchique de la chaîne d'analyse introduit dans Ch. 3, nous identifions les paramètres à estimer, écrivons la loi jointe et déterminons l'expression de l'estimateur pour préparer la mise en œuvre de l'algorithme.

À travers la modélisation directe hiérarchique de l'acquisition d'une donnée Sect. 3.5, nous avons introduit tous les paramètres nécessaires pour la quantification de protéines à base de spectrometrie de masse. Cette modélisation hiérarchique profite de la notion de dépendance et d'indépendance entre paramètres que l'on retrouve dans la modélisation probabiliste et dans les écritures de loi qui suivront.

4.2.1 Paramètres à estimer

Le but de la protéomique quantitative est d'estimer le plus précisément possible la quantité d'une molécule ciblée. Ainsi, dans cette thèse, le paramètre d'intérêt de la quantification est la concentration des protéines ciblées $\boldsymbol{x}$.

2. En effet, dans l'exemple du lien *protéines* $\rightarrow$ *peptides*, en choisissant la précision $\boldsymbol{\Gamma}_{\boldsymbol{\kappa}}$ infiniment grande, la loi de probabilité devient un Dirac au point $\mathbf{D}\boldsymbol{x}$, et nous avons forcé un lien déterministe.

La modélisation fait intervenir d'autres paramètres qui sont nécessaires pour l'explication des données, notamment la quantité peptidique $\boldsymbol{\kappa}$ et les paramètres instruments $\boldsymbol{\theta}^{\text{inst}}$ incluant les paramètres chromatographiques, les paramètres de bruit et les gains divers. Rappelons que $\boldsymbol{\theta}^{\text{inst}}$ est attaché à l'instrument utilisé et varie selon l'utilisation du mode Full-MS ou SRM. Le groupement de ces paramètres $(\boldsymbol{\kappa}, \boldsymbol{\theta}^{\text{inst}})$ est appelé *paramètres de nuisance* [3, Ch. 3]. On peut choisir de ne pas donner d'estimation de ces paramètres (puisqu'ils ne sont pas des « paramètres d'intérêt » au sens de l'inversion) et de les marginaliser dans les expressions probabilistes. C'est ce que nous ferons quand nous parlerons de classification (Ch. 5). Lors de la mise en œuvre *via* un échantillonnage stochastique de type MCMC cependant, nous devons tirer des échantillons aléatoires pour ces paramètres de nuisance. Leur estimation, faite à partir des échantillons stochastiques, est alors un sous-produit gratuit de l'estimation du paramètre d'intérêt par MCMC, même si ces paramètres sont marginalisés dans l'estimation de la concentration. Dans ce chapitre finalement, nous estimerons également les paramètres de nuisance.

4.2.2 Loi jointe

Compte tenu de la modélisation et de la hiérarchie, schématisé dans la Fig. 3.12, la loi jointe s'écrit

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\text{inst}}, \mathbf{Y}) = \mathrm{p}(\mathbf{Y} \mid \boldsymbol{\kappa}, \boldsymbol{\theta}^{\text{inst}}) \cdot \mathrm{p}(\boldsymbol{\kappa} \mid \boldsymbol{x})\, \mathrm{p}(\boldsymbol{\theta}^{\text{inst}}) \cdot \mathrm{p}(\boldsymbol{x}). \tag{4.1}$$

Les distributions pour ces lois ont été choisies et leur choix justifié dans le Ch. 3.

La loi jointe explique conjointement les paramètres et les données et marque le point de départ de nos calculs. Toutes les autres lois se déduisent d'elle par marginalisation de paramètres comme notamment la loi *a posteriori* et les lois *a posteriori* conditionnelles dont nous donnerons les calculs détailles en annexe de ce document (Ann. B).

Remarque ⟨**4.3**⟩ *La loi a priori de la concentration des protéines peut bénéficier d'une connaissance préalable sur les hyperparamètres de la distribution, ou elle peut l'ignorer.*

Dans le premier cas, la distribution de la concentration est connue, grâce à des études antérieures. C'est par exemple le cas quand l'individu dont provient l'échantillon biologique est atteint d'une maladie spécifique dont la distribution a déjà été étudiée. Un autre exemple est celui des échantillons biologiques synthétiques où une valeur nominale est connue (moyenne) avec une incertitude donnée (précision).

Dans le deuxième cas, quasiment aucune information a priori ne peut être donnée. Pour pouvoir attribuer une loi qui ne favorise pas de valeurs, nous approchons une loi uniforme sur l'ensemble $\mathbb{R}$ en faisant tendre la précision vers $\varepsilon > 0$ (voir Sect. 2.4). Ceci est notamment utilisé quand on ne connaît pas la maladie (ou classe) de provenance de l'échantillon biologique. Notons cependant que théoriquement, des valeurs négatives sont possibles alors que la concentration est non négative. Ceci est à surveiller dans l'analyse des résultats, issus de l'inversion avec cette loi a priori.

L'inférence sur les paramètres d'intérêt et de nuisance se fait à partir de la loi *a posteriori* totale, *i.e.* la loi de probabilité de tous les paramètres sachant les mesures. Par la règle de Bayes, elle est proportionnelle à la loi jointe avec le facteur de normalisation $\mathrm{p}(\mathbf{Y})$:

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\text{inst}} \mid \mathbf{Y}) = \frac{\mathrm{p}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\text{inst}}, \mathbf{Y})}{\mathrm{p}(\mathbf{Y})}. \tag{4.2}$$

Détails pour le cas LC-Full-MS

La loi jointe dans le cas de la modélisation du couplage LC-Full-MS s'écrivant

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\xi}, \gamma, \mathbf{Y}) = \mathrm{p}(\boldsymbol{x}) \cdot \mathrm{p}(\boldsymbol{\kappa} \mid \boldsymbol{x})\, \mathrm{p}(\boldsymbol{\tau})\, \mathrm{p}(\boldsymbol{\xi})\, \mathrm{p}(\gamma) \cdot \mathrm{p}(\mathbf{Y} \mid \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\xi}, \gamma),$$

la loi *a posteriori* totale a l'expression suivante :

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\xi}, \boldsymbol{\tau}, \gamma \mid \mathbf{Y}) = \frac{\mathrm{p}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\xi}, \boldsymbol{\tau}, \gamma, \mathbf{Y})}{\mathrm{p}(\mathbf{Y})}. \tag{4.3}$$

Détails pour le cas LC-SRM

Dans le cas d'utilisation du couplage LC-SRM, la loi jointe se décompose de la manière suivante :

$$\begin{aligned}\mathrm{p}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \gamma, \gamma^*, \boldsymbol{y}_{1:L}, \boldsymbol{y}^*_{1:L}) \\ = \mathrm{p}(\boldsymbol{x}) \cdot \mathrm{p}(\boldsymbol{\kappa} \mid \boldsymbol{x})\, \mathrm{p}(\boldsymbol{\tau})\, \mathrm{p}(\boldsymbol{\lambda})\, \mathrm{p}(\boldsymbol{\beta})\, \mathrm{p}(\boldsymbol{\phi}^*)\, \mathrm{p}(\gamma)\, \mathrm{p}(\gamma^*) \cdot \\ \mathrm{p}(\boldsymbol{y}_{1:L} \mid \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \gamma)\, \mathrm{p}(\boldsymbol{y}^*_{1:L} \mid \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \gamma^*).\end{aligned}$$

La loi *a posteriori* totale s'écrit donc

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \gamma, \gamma^* \mid \boldsymbol{y}_{1:L}, \boldsymbol{y}^*_{1:L}) = \frac{\mathrm{p}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \gamma, \gamma^*, \boldsymbol{y}_{1:L}, \boldsymbol{y}^*_{1:L})}{\mathrm{p}(\boldsymbol{y}_{1:L}, \boldsymbol{y}^*_{1:L})}. \tag{4.4}$$

4.2.3 Expression de l'estimateur

Le but de l'inversion est de quantifier les paramètres en utilisant un *quantifieur* noté $\psi^{\mathrm{q}} : \Omega_{\mathbf{Y}} \to \Omega_{\boldsymbol{x},\boldsymbol{\kappa},\boldsymbol{\theta}^{\mathrm{inst}}}$, la quantification étant une estimation ponctuelle de paramètre continu. Le choix de l'estimateur est attaché à une fonction de coût, comme nous l'avons introduit dans le Ch. 2. Nous choisissons ici le coût quadratique, pénalisant plus fortement les grands écarts que les petits écarts de la quantité vraie. L'estimateur qui y est associé est l'estimateur Espérance A Posteriori dont l'expression est

$$\psi^{\mathrm{q}}(\mathbf{Y}) = \int_{\Omega_{\boldsymbol{x},\boldsymbol{\kappa},\boldsymbol{\theta}^{\mathrm{inst}}}} \left[\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}}\right] \cdot \mathrm{p}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}} \mid \mathbf{Y})\, \mathrm{d}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}}). \tag{4.5}$$

L'expression précédente fait intervenir un produit de lois aboutissant à une loi non standard nécessitant l'intégration d'un grand nombre de données et de paramètres. Ainsi, le calcul du premier moment de la loi *a posteriori* n'est pas faisable analytiquement. Nous pouvons transformer le problème d'intégration des paramètres par rapport à une mesure de probabilité en un problème d'échantillonnage stochastique, voir Sect. 2.9.1.

À partir d'un ensemble d'échantillons stochastiques des paramètres, nous pouvons calculer le quantifieur en les moyennant à partir d'un certain indice $K_0 \in \mathbb{N}_0$:

$$\left[\hat{\boldsymbol{x}}, \hat{\boldsymbol{\kappa}}, \hat{\theta}^{\mathrm{inst}}\right] = \frac{1}{K} \sum_{k=K_0+1}^{K_0+K} \left[\boldsymbol{x}^{(k)}, \boldsymbol{\kappa}^{(k)}, (\boldsymbol{\theta}^{\mathrm{inst}})^{(k)}\right]. \tag{4.6}$$

En général, les moments d'ordre p sont obtenus en moyennant les échantillons élevés à la puissance p :

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{\kappa},\boldsymbol{\theta}^{\mathrm{inst}}|\mathbf{Y}} \left(\left[\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}}\right]^{[p]} \right) \approx \frac{1}{K} \sum_{k=K_0+1}^{K_0+K} \left[\boldsymbol{x}^{(k)}, \boldsymbol{\kappa}^{(k)}, (\boldsymbol{\theta}^{\mathrm{inst}})^{(k)}\right]^{[p]}$$

avec $[x_1, \ldots, x_n]^{[p]} := [x_1^p, \ldots, x_n^p]$. Ceci nous permettra de calculer également la variance de la loi *a posteriori*, donnée par la différence du carré du moment d'ordre 1 et du moment d'ordre 2, pour déterminer l'intervalle de crédibilité de l'estimation.

Le tirage des échantillons stochastiques par l'utilisation des méthodes MCMC fait objet de la section suivante.

4.3 Mise en œuvre de l'inversion

Nous voulons calculer numériquement l'intégrale de l'Éqn. (4.5) du fait de l'impossibilité du calcul analytique. Pour cela, nous proposons d'utiliser l'approche MCMC en adoptant une structure de Gibbs (Sect. 2.9.1). Elle nous permet de calculer le quantifieur en moyennant les échantillons tirés sous la loi *a posteriori* jointe des paramètres, comme exprimé dans l'Éqn. (4.6).

Pour mettre en place la structure de Gibbs, il nous faut (1) identifier la structure qui permet d'échantillonner le plus efficacement par regroupement de paramètres, (2) donner les expressions des lois *a posteriori* conditionnelles, puis (3) décider de l'algorithme afin de générer les échantillons de ces dernières.

Quant au premier point et compte tenu de l'attribution des lois *a priori*, nous proposons d'échantillonner paramètre par paramètre (vectoriel quand cela est possible). Étudions à présent les lois *a posteriori* conditionnelles et les échantillonneurs associés.

4.3.1 Lois a posteriori conditionnelles

Nous allons passer en revue tous les paramètres et lister les propriétés nécessaires pour déterminer la loi *a posteriori* conditionnelle de chaque paramètre. Les calculs et les expressions des hyperparamètres sont explicités dans l'Ann. B.

Concentration de protéines

Nous recensons d'abord la loi *a posteriori* conditionnelle de la concentration des protéines. Elle est commune aux deux modes de spectrométrie. Comme précédemment, nous utilisons la convention $\Delta = \mathbf{D}$ dans le cas LC-Full-MS et $\Delta = \boldsymbol{\chi}\mathbf{D}\boldsymbol{\psi}$ dans le cas SRM.

(1) concentration des protéines cibles $\boldsymbol{x}$

<u>a priori :</u> $\mathrm{p}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_x, \boldsymbol{\Gamma}_x)$
<u>vraisemblance associée :</u> $\mathrm{p}(\boldsymbol{\kappa} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{\kappa}; \Delta\boldsymbol{x}, \boldsymbol{\Gamma}_\kappa)$
<u>conjugaison :</u> oui car la vraisemblance pour $\boldsymbol{x}$ a la même forme que son *a priori*
<u>a posteriori :</u> $\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\kappa}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_x^{\mathrm{post}}, \boldsymbol{\Gamma}_x^{\mathrm{post}})$
- $\boldsymbol{m}_x^{\mathrm{post}} = (\boldsymbol{\Gamma}_x^{\mathrm{post}})^{-1}(\boldsymbol{\Gamma}_x \boldsymbol{m}_x + \Delta^{\mathrm{T}}\boldsymbol{\Gamma}_\kappa \boldsymbol{\kappa})$,
- $\boldsymbol{\Gamma}_x^{\mathrm{post}} = \boldsymbol{\Gamma}_x + \Delta^{\mathrm{T}}\boldsymbol{\Gamma}_\kappa \Delta$,

<u>échantillonnage :</u> explicite

Détails pour le cas LC-Full-MS

($2^{\text{Full-MS}}$) quantité de peptides $\boldsymbol{\kappa}$

<u>a priori :</u> $\mathrm{p}(\boldsymbol{\kappa} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{\kappa}; \Delta\boldsymbol{x}, \boldsymbol{\Gamma}_\kappa)$
<u>vraisemblance associée :</u> $\mathrm{p}(\mathbf{Y} \mid \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}}) = \mathcal{N}(\mathbf{Y}; \mathbf{M} + \mathbf{M}^*, \gamma)$ où $\mathbf{M}$ peut être réécrit pour la variable $\boldsymbol{\kappa}$ comme $\mathbf{M} = \mathbf{H}\boldsymbol{\kappa}$ (Sect. 3.4.1)
<u>conjugaison :</u> oui car la vraisemblance pour $\boldsymbol{\kappa}$ a la même forme que son *a priori*
<u>a posteriori :</u> $\mathrm{p}(\boldsymbol{\kappa} \mid \boldsymbol{x}, \boldsymbol{\kappa}^*, \boldsymbol{\xi}, \boldsymbol{\tau}, \gamma, \mathbf{Y}) = \mathcal{N}(\boldsymbol{\kappa}; \boldsymbol{m}_\kappa^{\mathrm{post}}, \boldsymbol{\Gamma}_\kappa^{\mathrm{post}})$
- $\boldsymbol{m}_\kappa^{\mathrm{post}} = (\boldsymbol{\Gamma}_\kappa^{\mathrm{post}})^{-1}\left(\boldsymbol{\Gamma}_\kappa \Delta\boldsymbol{x} + \mathbf{H}^{\mathrm{T}}\boldsymbol{\Gamma}_{\mathbf{Y}}\boldsymbol{y} + (\mathbf{H}^*)^{\mathrm{T}}\boldsymbol{\Gamma}_{\mathbf{Y}}\mathbf{H}^*\boldsymbol{\kappa}^*\right)$,
- $\boldsymbol{\Gamma}_\kappa^{\mathrm{post}} = \boldsymbol{\Gamma}_\kappa + \mathbf{H}^{\mathrm{T}}\boldsymbol{\Gamma}_{\mathbf{Y}}\mathbf{H}$,

<u>échantillonnage :</u> explicite

($3^{\text{Full-MS}}$) gain de système $\boldsymbol{\xi}$

<u>a priori :</u> $\mathrm{p}(\boldsymbol{\xi}) = \mathcal{N}(\boldsymbol{\xi}; \boldsymbol{m}_\xi, \boldsymbol{\Gamma}_\xi)$

vraisemblance associée : $\mathrm{p}(\mathbf{Y} \mid \boldsymbol{\kappa}, \boldsymbol{\kappa}^*, \boldsymbol{\xi}, \boldsymbol{\tau}, \gamma) = \mathcal{N}(\mathbf{Y}; \mathbf{M} + \mathbf{M}^*, \gamma)$ où la somme $\mathbf{M} + \mathbf{M}^*$ peut être réécrite pour la variable $\boldsymbol{\xi}$ comme $\mathbf{M} + \mathbf{M}^* = \mathbf{G}\boldsymbol{\xi}$, voir Sect. 3.4.1

conjugaison : oui car la vraisemblance pour $\boldsymbol{\xi}$ a la même forme que son *a priori*

a posteriori : $\mathrm{p}(\boldsymbol{\xi} \mid \boldsymbol{\kappa}, \boldsymbol{\kappa}^*, \boldsymbol{\tau}, \gamma, \mathbf{Y}) = \mathcal{N}(\boldsymbol{\xi}; \boldsymbol{m}_{\xi}^{\mathrm{post}}, \boldsymbol{\Gamma}_{\xi}^{\mathrm{post}})$

- $\boldsymbol{m}_{\xi}^{\mathrm{post}} = (\boldsymbol{\Gamma}_{\xi}^{\mathrm{post}})^{-1} \left(\boldsymbol{\Gamma}_{\xi} \boldsymbol{m}_{\xi} + \mathbf{G}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \boldsymbol{y}\right)$,
- $\boldsymbol{\Gamma}_{\xi}^{\mathrm{post}} = \boldsymbol{\Gamma}_{\xi} + \mathbf{G}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{G}$,

échantillonnage : explicite

($4^{\text{Full-MS}}$) temps de rétention chromatographique $\boldsymbol{\tau}$

a priori : $\mathrm{p}(\boldsymbol{\tau}) = \mathcal{U}(\boldsymbol{\tau}; \boldsymbol{\tau}^{\mathrm{m}}, \boldsymbol{\tau}^{\mathrm{M}})$

vraisemblance associée : $\mathrm{p}(\mathbf{Y} \mid \boldsymbol{\kappa}, \boldsymbol{\kappa}^*, \boldsymbol{\xi}, \boldsymbol{\tau}, \gamma) = \mathcal{N}(\mathbf{Y}; \mathbf{M} + \mathbf{M}^*, \gamma)$

conjugaison : non car la vraisemblance pour $\boldsymbol{\tau}$ n'a pas de forme usuelle ([1, Ann. 7.4]) qui ne conjugue aucune loi *a priori* standard

a posteriori : $\mathrm{p}(\boldsymbol{\tau} \mid \boldsymbol{\kappa}, \boldsymbol{\kappa}^*, \boldsymbol{\xi}, \gamma, \mathbf{Y}) \propto \mathrm{p}(\boldsymbol{\tau}) \cdot \mathrm{p}(\mathbf{Y} \mid \boldsymbol{\kappa}, \boldsymbol{\kappa}^*, \boldsymbol{\xi}, \boldsymbol{\tau}, \gamma) = \mathcal{U}(\boldsymbol{\tau}; \boldsymbol{\tau}^{\mathrm{m}}, \boldsymbol{\tau}^{\mathrm{M}}) \cdot \mathcal{N}(\mathbf{Y}; \mathbf{M} + \mathbf{M}^*, \gamma)$

échantillonnage : une itération de Metropolis-Hasting à Marche Aléatoire

($5^{\text{Full-MS}}$) inverse variance du bruit γ

a priori : $\mathrm{p}(\gamma) = \mathcal{G}(\gamma; \alpha_{\gamma}, \beta_{\gamma})$

vraisemblance associée : $\mathrm{p}(\mathbf{Y} \mid \boldsymbol{\kappa}, \boldsymbol{\kappa}^*, \boldsymbol{\xi}, \boldsymbol{\tau}, \gamma) = \mathcal{N}(\mathbf{Y}; \mathbf{M} + \mathbf{M}^*, \gamma)$

conjugaison : oui car la vraisemblance pour γ a la même forme que son *a priori*

a posteriori : $\mathrm{p}(\gamma \mid \boldsymbol{\kappa}, \boldsymbol{\kappa}^*, \boldsymbol{\xi}, \boldsymbol{\tau}, \mathbf{Y}) = \mathcal{G}(\gamma; \alpha_{\gamma}^{\mathrm{post}}, \beta_{\gamma}^{\mathrm{post}})$

- $\alpha^{\mathrm{post}} = N/2$,
- $\beta^{\mathrm{post}} = 2/\left\|\mathbf{Y} - \mathbf{M} - \mathbf{M}^*\right\|^2$,

échantillonnage : explicite

Remarque ⟨4.4⟩ (Calcul de matrices) *Les calculs des produits de matrices sont très gourmands dû à la dimension des données. Dans [1], les produits ont été décomposés en des facteurs de dimension beaucoup plus petite. Ainsi, nous ne calculons pas par exemple la matrice* $\mathbf{H}$ *pour la multiplier avec sa transposée, mais déduisons directement l'expression explicite pour* $\mathbf{H}^{\mathrm{T}}\mathbf{H}$. *Idem pour* $\mathbf{G}^{\mathrm{T}}\mathbf{G}$, $\mathbf{H}^{\mathrm{T}}\mathbf{Y}$, *. . . . Nous les utiliserons pour accélerer les calculs et pour utiliser moins de mémoire.*

Détails pour le cas LC-SRM

Pour le cas LC-SRM que nous considérons dans ce paragraphe, soit $\mathcal{L}_i$ l'ensemble des indices $l = 1, \ldots, L$ tels que la trace l, associée au fragment l, soit issue du peptide i.

(2^{SRM}) quantité de peptides $\boldsymbol{\kappa}$

a priori : $\mathrm{p}(\boldsymbol{\kappa} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{\kappa}; \Delta\boldsymbol{x}, \boldsymbol{\Gamma}_{\boldsymbol{\kappa}}) = \prod_{i=1}^{I} \mathcal{N}(\kappa_i; \sum_p \delta_{ip} x_p, \gamma_{\kappa_i})$

vraisemblance associée : $\mathrm{p}(\boldsymbol{y}_{1:L} \mid \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_i \prod_{l \in \mathcal{L}} \mathrm{p}(\boldsymbol{y}_l \mid \kappa_i, \tau_i, \lambda_i, \beta_l, \gamma_l)$

conjugaison : oui car la vraisemblance pour $\boldsymbol{\kappa}$ a la même forme que son *a priori*

a posteriori : $\mathrm{p}(\boldsymbol{\kappa} \mid \boldsymbol{y}_{1:L}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{x}) = \mathcal{N}(\boldsymbol{\kappa}; \boldsymbol{m}_{\boldsymbol{\kappa}}^{\mathrm{post}}, \boldsymbol{\Gamma}_{\boldsymbol{\kappa}}^{\mathrm{post}}) = \prod_{i=1}^{I} \mathcal{N}(\kappa_i; m_{\kappa_i}{}^{\mathrm{post}}, \gamma_{\kappa_i}^{\mathrm{post}})$

- $m_{\kappa_i}^{\mathrm{post}} = (\gamma_{\kappa_i}^{\mathrm{post}})^{-1} \left[\gamma_{\kappa_i} \sum_p \delta_{ip} \boldsymbol{x}_p + \sum_{l \in \mathcal{L}_i} \gamma_l \beta_l \mathscr{C}_i^{\mathrm{T}} \boldsymbol{y}_l\right]$
- $\gamma_{\kappa_i}^{\mathrm{post}} = \gamma_{\kappa_i} + \sum_{l \in \mathcal{L}_i} \gamma_l \beta_l^2 \mathscr{C}_i^{\mathrm{T}} \mathscr{C}_i$

échantillonnage : explicite

(3^{SRM}) positions des pics chromatographiques $\boldsymbol{\tau}$

a priori : $\mathrm{p}(\boldsymbol{\tau}) = \mathcal{U}(\boldsymbol{\tau}; \boldsymbol{\tau}^{\mathrm{m}}, \boldsymbol{\tau}^{\mathrm{M}})$

vraisemblance associée : $\mathrm{p}(\boldsymbol{y}_{1:L}, \boldsymbol{y}_{1:L}^* \mid \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \gamma, \gamma^*)$

conjugaison : non, cas similaire à [1, Ann. 7.4] : conjugaison avec aucune loi impliquée
a posteriori : $\mathrm{p}(\boldsymbol{\tau} \,|\, \boldsymbol{y}_{1:L}, \boldsymbol{y}^*_{1:L}, \boldsymbol{\kappa}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \boldsymbol{\gamma}, \boldsymbol{\gamma}^*) \propto \mathrm{p}(\boldsymbol{\tau})\, \mathrm{p}(\boldsymbol{y}_{1:L}, \boldsymbol{y}^*_{1:L} \,|\, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$
échantillonnage : une itération de Metropolis-Hastings à Marche Aléatoire

(4^{SRM}) largeurs des pics chromatographics $\boldsymbol{\lambda}$
a priori : $\mathrm{p}(\boldsymbol{\lambda}) = \mathcal{U}(\boldsymbol{\tau}; \left[\lambda^{\mathrm{m}}, \lambda^{\mathrm{M}}\right]^I)$
vraisemblance associée : $\mathrm{p}(\boldsymbol{y}_{1:L}, \boldsymbol{y}^*_{1:L} \,|\, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$
conjugaison : non, cas similaire à [1, Ann. 7.4] : conjugaison avec aucune loi impliquée
a posteriori : $\mathrm{p}(\boldsymbol{\lambda} \,|\, \boldsymbol{y}_{1:L}, \boldsymbol{y}^*_{1:L}, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \boldsymbol{\gamma}, \boldsymbol{\gamma}^*) \propto \mathrm{p}(\boldsymbol{\lambda})\, \mathrm{p}(\boldsymbol{y}_{1:L}, \boldsymbol{y}^*_{1:L} \,|\, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$
échantillonnage : une itération de Metropolis-Hastings à Marche Aléatoire

(5^{SRM}) gains de fragmentation $\boldsymbol{\beta}$
a priori : $\mathrm{p}(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{m}_{\boldsymbol{\beta}}, \boldsymbol{\Gamma}_{\boldsymbol{\beta}}) = \prod_l \mathcal{N}(\beta_l; m_{\beta_l}, \gamma_{\beta_l})$
vraisemblance associée : $\mathrm{p}(\boldsymbol{y}_{1:L}, \boldsymbol{y}^*_{1:L} \,|\, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$
conjugaison : oui car la vraisemblance pour $\boldsymbol{\beta}$ a la même forme que son *a priori*
a posteriori : $\mathrm{p}(\boldsymbol{\beta} \mid \boldsymbol{y}_{1:L}, \boldsymbol{y}^*_{1:L}, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\phi}^*, \boldsymbol{\gamma}, \boldsymbol{\gamma}^*) = \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{m}^{\mathrm{post}}_{\boldsymbol{\beta}}, \boldsymbol{\Gamma}^{\mathrm{post}}_{\boldsymbol{\beta}}) = \prod_l \mathcal{N}(\beta_l; m^{\mathrm{post}}_{\beta_l}, \gamma^{\mathrm{post}}_{\beta_l})$
- $m_{\beta_l} = (\gamma^{\mathrm{post}}_{\beta_l})^{-1} \left[\gamma_{\beta_l} + \gamma_l \kappa_i \mathscr{C}_i^{\mathrm{T}} \boldsymbol{y}_l + \gamma^*_l \phi^*_l \kappa^*_i \mathscr{C}_i^{\mathrm{T}} \boldsymbol{y}^*_l\right]$
- $\gamma_{\beta_l} = \gamma_{\beta_l} + \left(\gamma_l \kappa_i^2 + \gamma^*_l \phi^*_l \kappa^*_i\right) \mathscr{C}_i^{\mathrm{T}} \mathscr{C}_i$

échantillonnage : explicite

(6^{SRM}) gains de transition $\boldsymbol{\phi}^*$
a priori : $\mathrm{p}(\boldsymbol{\phi}^*) = \prod_l \mathrm{p}(\phi^*_l) = \prod_l \mathcal{N}(\phi^*_l; 1, \gamma_{\phi^*})$
vraisemblance associée : $\mathrm{p}(\boldsymbol{y}^*_{1:L} \,|\, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \boldsymbol{\gamma}^*)$
conjugaison : oui car la vraisemblance pour $\boldsymbol{\phi}^*$ a la même forme que son *a priori*
a posteriori : $\mathrm{p}(\boldsymbol{\phi}^* \,|\, \boldsymbol{y}^*_{1:L}, \boldsymbol{\beta}, \boldsymbol{\gamma}^*, \boldsymbol{\tau}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{\phi}^*; \boldsymbol{m}^{\mathrm{post}}_{\boldsymbol{\phi}^*}, \boldsymbol{\Gamma}^{\mathrm{post}}_{\boldsymbol{\phi}^*}) = \prod_l \mathcal{N}(\phi^*_l; m^{\mathrm{post}}_{\phi^*_l}, \gamma^{\mathrm{post}}_{\phi^*_l})$
- $m^{\mathrm{post}}_{\phi^*_l} = \left(\gamma^{\mathrm{post}}_{\phi^*_l}\right)^{-1} \left[\gamma_{\phi^*_l} + \gamma^*_l \kappa^*_i \beta_l \mathscr{C}_i^{\mathrm{T}} \boldsymbol{y}^*_l\right]$
- $\gamma^{\mathrm{post}}_{\phi^*_l} = \gamma_{\phi^*_l} + \gamma^*_l (\kappa^*_i \beta_l)^2 \mathscr{C}_i^{\mathrm{T}} \mathscr{C}_i$

échantillonnage : explicite

(7^{SRM}) inverse variance du bruit de mesure $\boldsymbol{\gamma}$
a priori : $\mathrm{p}(\boldsymbol{\gamma}) = \prod_l \mathrm{p}(\gamma_l) = \prod_l \mathcal{G}(\gamma_l; \alpha_l, \beta_l)$
vraisemblance associée : $\mathrm{p}(\boldsymbol{y}_{1:L} \,|\, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_l \mathrm{p}(\boldsymbol{y}_l \,|\, \kappa_i, \tau_i, \lambda_i, \beta_l, \gamma_l)$
conjugaison : oui car pour chaque l la vraisemblance pour γ_l a la même forme que son *a priori*
a posteriori : $\mathrm{p}(\boldsymbol{\gamma} \,|\, \boldsymbol{y}_{1:L}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\kappa}) = \prod_l \mathrm{p}(\gamma_l \,|\, \boldsymbol{y}_l, \beta_l, \tau_i, \lambda_i, \kappa_i) = \prod_l \mathcal{G}(\gamma_l; \alpha^{\mathrm{post}}_l, \beta^{\mathrm{post}}_l)$
- $\alpha^{\mathrm{post}}_l = \alpha_l + N_l/2,$
- $\beta^{\mathrm{post}}_l = \left[\beta_l^{-1} + \left\|\boldsymbol{y}_l - \beta_l \kappa_i \mathscr{C}_i\right\|^2/2\right]^{-1}$

échantillonnage : explicite

(8^{SRM}) inverse variance du bruit de mesure marquée $\boldsymbol{\gamma}^*$
a priori : $\mathrm{p}(\boldsymbol{\gamma}^*) = \prod_l \mathrm{p}(\gamma^*_l) = \prod_l \mathcal{G}(\gamma^*_l; \alpha^*_l, \beta^*_l)$
vraisemblance associée : $\mathrm{p}(\boldsymbol{y}^*_{1:L} \,|\, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \boldsymbol{\gamma}^*) = \prod_l \mathrm{p}(\boldsymbol{y}^*_l \,|\, \tau_i, \lambda_i, \beta_l, \phi^*_l, \gamma^*_l)$
conjugaison : oui car pour chaque l la vraisemblance pour γ^*_l a la même forme que son *a priori*
a posteriori : $\mathrm{p}(\boldsymbol{\gamma}^* \mid \boldsymbol{y}^*_{1:L}, \boldsymbol{\beta}, \boldsymbol{\phi}^*, \boldsymbol{\tau}, \boldsymbol{\lambda}) = \prod_l \mathrm{p}(\gamma^*_l \mid \boldsymbol{y}^*_l, \beta_l, \phi^*_l, \tau_i, \lambda_i) = \prod_l \mathcal{G}(\gamma^*_l; (\alpha^*_l)^{\mathrm{post}}, (\beta^*_l)^{\mathrm{post}})$

- $(\alpha_l^*)^{\text{post}} = \alpha_l^* + N_l/2$,
- $(\beta_l^*)^{\text{post}} = \left[(\beta_l^*)^{-1} + \left\|\boldsymbol{y}_l^* - \beta_l \boldsymbol{\phi}_l^* \kappa_i^* \mathscr{C}_i\right\|^2 /2\right]^{-1}$

<u>échantillonnage :</u> explicite

4.3.2 Bilan

L'échantillonnage des paramètres dont les lois *a priori* sont conjuguées est facile puisque les lois *a posteriori* de ces paramètres sont connues sous forme usuelle. Pour celles-ci, il existe des générateurs d'échantillons explicites. Certains paramètres, cependant, n'ont pas de loi *a priori* conjuguée. C'est notamment le cas pour les paramètres chromatographiques où nous avons choisi une loi uniforme multidimensionnelle (voir [1, Sect. 4.3.3.3]). Pour l'échantillonnage sous la loi *a posteriori* conditionnelle comme loi cible, nous avons recours à l'algorithme de Metropolis-Hastings (Sect. 2.9.1). Nous choisissons d'utiliser une marche aléatoire avec un noyau gaussien comme loi de proposition. En faisant ce choix, nous avons un bon compromis entre temps de calcul et convergence : les propositions de valeurs ne nécessitent qu'un tirage aléatoire sous une densité normale. La marche aléatoire permet à la chaîne de converger en un temps raisonnable vers la zone de forte probabilité. Rappelons que l'algorithme de Metropolis-Hastings à Marche Aléatoire a un paramètre de définition de voisinage qui est à configurer pour chaque application.

Le pseudo-code de l'algorithme est résumé dans Alg. 4.1, page 79.

4.4 Étalonnage par Contrôlé de Qualité

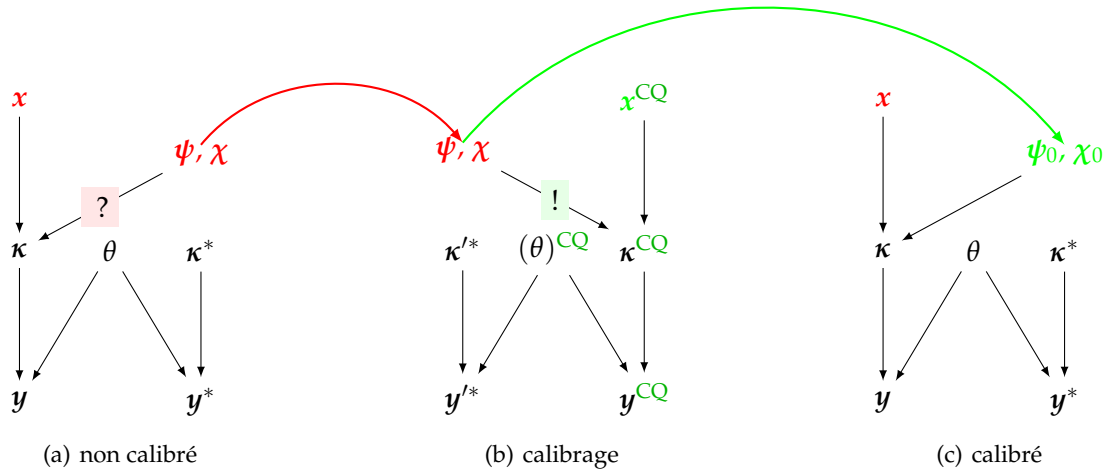
Les acquisitions de données en mode SRM fournies par notre partenaire **bioMérieux** dans le cadre du projet ANR BHI-PRO contiennent un étalonnage interne au niveau des peptides. Grâce au standard AQUA (voir Sect. 3.1.4), nous gérons les variabilités et gains jusqu'au niveau hiérarchique des peptides. Cependant, le lien entre peptides et protéines n'est pas traçable dans une seule acquisition avec cet étalonnage. Ce lien comporte notamment le gain de préparation $\boldsymbol{\psi}$ et celui de digestion $\boldsymbol{\chi}$. Si on suppose stables les variabilités d'un même jour concernant la préparation et la digestion, il suffit d'un calibrage qui permettrait la réutilisation dans le cadre de l'inversion. Pour cette procédure, la concentration protéique ne doit plus être inconnue.

En effet, les campagnes expérimentales de bioMérieux sont « encadrées » par des passages d'échantillons synthétiques préparés sous les mêmes conditions que les échantillons cliniques. Il s'agit des échantillons de *Contrôle de Qualité*. Ils sont constitués du même mélange complexe que les échantillons d'intérêt et enrichis de telle manière que les concentrations protéiques résultantes sont — à des erreurs de synthèse près — toutes égales à une valeur donnée. C'est à l'aide de ces données que nous reconstruirons le lien entre protéines et peptides par un calibrage externe, le résultat de ce dernier étant injecté ensuite dans l'algorithme de quantification [3], illustré dans la Fig. 4.2.

Pour simplifier le problème, nous réécrivons le produit des gains et de la matrice de

3. et de classification ...

Fig. 4.2 Illustration de la nécessité de l'étalonnage par CQ. Le paramètre θ regroupe les paramètres partagés.
(a) Les inconnus sont la concentration et les gains. Le calibrage peut être reporté aux échantillons synthétiques de la même campagne de préparation.
(b) Les échantillons étant synthétisés, la concentration est connue. La seule inconnue est le couple de gains qui est calibré.
(c) Le scénario de (a) est modifié : la seule inconnue restante est la concentration, les gains étant calibrés par les échantillons CQ.



digestion $\chi\mathbf{D}\boldsymbol{\psi}$:

$$
\begin{aligned}
\chi\mathbf{D}\boldsymbol{\psi} &= \begin{bmatrix} \chi_1 & & \\ & \ddots & \\ & & \chi_I \end{bmatrix} \begin{bmatrix} d_{11} & \dots & d_{I,1} \\ \vdots & & \vdots \\ d_{1,P} & \dots & d_{I,P} \end{bmatrix} \begin{bmatrix} \psi_1 & & \\ & \ddots & \\ & & \psi_P \end{bmatrix} \\
&= \begin{bmatrix} \chi_1 d_{11}\psi_1 & \dots & \chi_I d_{I,1}\psi_1 \\ \vdots & & \vdots \\ \chi_1 d_{1,P}\psi_P & \dots & \chi_I d_{I,P}\psi_P \end{bmatrix} \\
&= \begin{bmatrix} \chi_1\psi_1 & \dots & \chi_I\psi_1 \\ \vdots & & \vdots \\ \chi_1\psi_P & \dots & \chi_I\psi_P \end{bmatrix} \odot \begin{bmatrix} d_{11} & \dots & d_{I,1} \\ \vdots & & \vdots \\ d_{1,P} & \dots & d_{I,P} \end{bmatrix} \\
&= \mathbf{G} \odot \mathbf{D}
\end{aligned}
$$

où $\odot$ désigne le produit matriciel de Hadamard, *i.e.* un produit terme à terme. La matrice $\mathbf{G} \in \mathbb{N}^{I\times P}$ est structurée par la matrice de digestion $\mathbf{D}$: elle ne prend des valeurs $[\mathbf{G}]_{ip} = g_{ip}$ que pour les indices (i, p) tels que $d_{ip} \neq 0$. Nous reportons donc le problème de calibrage des gains $\boldsymbol{\psi}$ et $\boldsymbol{\chi}$ au calibrage du produit des gains $\mathbf{G}$. La matrice étant parcimonieuse dans nos applications — on se souvient notamment de la protéotypie que nous supposons —, le calibrage sur elle revient à calibrer autant de valeurs qu'il y a de peptides.

Le but est de construire un estimateur de la matrice des gains à partir des données CQ. Si on note θ les paramètres de nuisance par rapport à l'estimation de $\mathbf{G}$ et que l'on choisit la fonction de coût quadratique $\mathrm{L}(\mathbf{G}, \psi_{\mathbf{G}}(\boldsymbol{y}^{\mathrm{CQ}})) = \left\|\mathbf{G} - \psi_{\mathbf{G}}(\boldsymbol{y}^{\mathrm{CQ}})\right\|_{\mathrm{F}}^2$, alors l'estimateur

est l'Espérance A Posteriori :

$$\psi(\boldsymbol{y}^{\mathrm{CQ}}) = \int_{\Omega_{\mathbf{G}}} \mathrm{p}(\mathbf{G} \mid \boldsymbol{y}^{CQ})\, \mathrm{d}\mathbf{G} = \int_{\Omega_{\mathbf{G}}} \int_{\Omega_{\theta}} \mathrm{p}(\mathbf{G}, \theta \mid \boldsymbol{y}^{CQ})\, \mathrm{d}\theta\, \mathrm{d}\mathbf{G}. \tag{4.7}$$

Pour l'estimation des g_{ip}, rappelons que l'expression de la quantité peptidique s'écrit

$$\kappa_i^{\mathrm{CQ}} = \sum_{p'=1}^{P'} g_{ip'} d_{ip'} x_{p'}^{\mathrm{CQ}} + \varepsilon_i = g_{ip} d_{ip} x_p^{\mathrm{CQ}} + \varepsilon_i$$

car il existe un seul p dont le peptide i est issue.

Le bruit additif étant modélisé gaussien et de moyenne nulle, nous écrivons la fonction de vraisemblance pour le paramètre g_{ip}

$$\mathrm{p}(\kappa_i^{\mathrm{CQ}} \mid g_{ip}) = \mathcal{N}\left(\kappa_i^{\mathrm{CQ}};\ g_{ip} d_{ip} x_p^{\mathrm{CQ}},\ \gamma_{\kappa_i}\right). \tag{4.8}$$

Pour tout couple $(i,p) \in \mathcal{I} := \{(i,p) | d_{ip} \neq 0\}$, nous choisissons une loi normale de moyenne m_g et de précision γ_g pour la loi *a priori* de g_{ip} : $\mathrm{p}(g_{ip}) = \mathcal{N}(g_{ip}; m_g, \gamma_g)$. Du fait de la faible connaissance des valeurs de ce paramètre, la distribution sera choisie vaguement informative, *i.e.* avec une précision très faible.

Nous pouvons écrire la loi jointe du modèle probabiliste de l'étalonnage utilisant les échantillons biologiques synthétiques CQ :

$$\begin{aligned}\mathrm{p}(\boldsymbol{y}^{\mathrm{CQ}}, \boldsymbol{y}^*, (\boldsymbol{\theta}^{\mathrm{inst}})^{\mathrm{CQ}}, \boldsymbol{\phi}^*, \boldsymbol{\kappa}^{\mathrm{CQ}}, \mathbf{G}) = \\ \mathrm{p}(\boldsymbol{y}^{\mathrm{CQ}} \mid (\boldsymbol{\theta}^{\mathrm{inst}})^{\mathrm{CQ}}, \boldsymbol{\kappa}^{\mathrm{CQ}})\, \mathrm{p}(\boldsymbol{y}^* \mid (\boldsymbol{\theta}^{\mathrm{inst}})^{\mathrm{CQ}}, \boldsymbol{\phi}^*)\, \mathrm{p}(\boldsymbol{\phi}^*)\, \mathrm{p}(\boldsymbol{\kappa}^{\mathrm{CQ}} \mid \mathbf{G})\, \mathrm{p}(\mathbf{G})\end{aligned} \tag{4.9}$$

avec $\mathrm{p}(\mathbf{G}) = \prod_{(i,p)\in\mathcal{I}} \mathrm{p}(g_{ip})$.

Afin de procéder à la marginalisation et à l'intégration, nous nous servons à nouveau des approches d'échantillonnage stochastique de type MCMC. Nous nous inspirons plus précisément de ce que nous avons déjà déduit pour la quantification. Il ne reste qu'une ligne à changer. Au lieu d'échantillonner le paramètre « concentration », désormais une quantité connue, nous échantillonnons le paramètre « matrice d'étalonnage CQ ». La loi *a priori* étant conjuguée, la loi *a posteriori* conditionnelle est également de forme normale de moyenne $m_{g,ip}^{\mathrm{post}}$ et de précision $\gamma_{g,ip}^{\mathrm{post}}$:

$$\begin{aligned}\gamma_{g,ip}^{\mathrm{post}} &= \gamma_{\kappa_i} d_{ip}^2 (x_p^{\mathrm{CQ}})^2 + \gamma_g, \\ m_{g,ip}^{\mathrm{post}} &= (\gamma_{g,ip}^{\mathrm{post}})^{-1} \left[\gamma_{\kappa_i} d_{ip} x_p^{\mathrm{CQ}} \kappa_i^{\mathrm{CQ}} + \gamma_g m_g.\right]\end{aligned}$$

Extension à l'utilisation de plusieurs échantillons CQ

Habituellement, à une date d'acquisition correspond (au moins) un échantillon synthétique de Contrôle de Qualité. Cependant, lors de la transmission des données, il arrive que les fichiers soient incomplets et que les données CQ d'un jour précis manquent. Pour intégrer malgré ceci l'effet de variabilité entre protéines et peptides, nous régularisons le problème par rapport à la date d'acquisition en moyennant sur tous les échantillons CQ disponibles. Cette approche demande la gestion d'un ensemble de données contenant le paramètre d'intérêt.

L'extension à l'utilisation de plusieurs échantillons synthétiques est immédiate. Soit N^{CQ} l'effectif de l'ensemble des échantillons CQ à disposition. Les acquisitions étant indépendantes modulo les paramètres stables, la quantité peptidique de chaque échantillon

est conditionnée par $\mathbf{G}$. En adoptant la notation $x_{1:N} = \{x_1, \ldots, x_N\}$, l'estimateur Espérance A Posteriori s'écrit

$$\psi(\boldsymbol{y}^{\mathrm{CQ}}_{1:N^{\mathrm{CQ}}}) = \int_{\Omega_{\mathbf{G}}} \mathrm{p}(\mathbf{G} \mid \boldsymbol{y}^{\mathrm{CQ}}_{1:N^{\mathrm{CQ}}})\, \mathrm{d}\mathbf{G} = \int_{\Omega_{\mathbf{G}}} \int_{\Omega_\theta} \mathrm{p}(\mathbf{G}, \theta_{1:N^{\mathrm{CQ}}} \mid \boldsymbol{y}^{\mathrm{CQ}}_{1:N^{\mathrm{CQ}}})\, \mathrm{d}\theta_{1:N^{\mathrm{CQ}}}\, \mathrm{d}\mathbf{G}. \tag{4.10}$$

La loi jointe du problème « multi-CQ » devient

$$\prod_{n=1}^{N^{\mathrm{CQ}}} \mathrm{p}(\boldsymbol{y}^{\mathrm{CQ}}_n, \boldsymbol{y}^*_n, (\boldsymbol{\theta}^{\mathrm{inst}})^{\mathrm{CQ}}_n, \boldsymbol{\phi}^*_n, \boldsymbol{\kappa}^{\mathrm{CQ}}_n, \mathbf{G}) = \prod_{n=1}^{N^{\mathrm{CQ}}} \left[\mathrm{p}(\boldsymbol{y}^{\mathrm{CQ}}_n \mid (\boldsymbol{\theta}^{\mathrm{inst}})^{\mathrm{CQ}}_n, \boldsymbol{\kappa}^{\mathrm{CQ}}_n)\, \mathrm{p}(\boldsymbol{y}^*_n \mid (\boldsymbol{\theta}^{\mathrm{inst}})^{\mathrm{CQ}}_n, \boldsymbol{\phi}^*_n)\, \mathrm{p}(\boldsymbol{\phi}^*_n\, \mathrm{p}(\boldsymbol{\kappa}^{\mathrm{CQ}}_n \mid \mathbf{G})\right] \mathrm{p}(\mathbf{G}) \tag{4.11}$$

et la loi *a posteriori* conditionnelle qui est toujours une loi normale a les paramètres suivants :

$$\gamma^{\mathrm{post}}_{g,ip} = \gamma_{\kappa_i} d^2_{ip} \sum_{n=1}^{N^{\mathrm{CQ}}} (x^{\mathrm{CQ}}_{p;\,n})^2 + \gamma_g,$$

$$m^{\mathrm{post}}_{g,ip} = (\gamma^{\mathrm{post}}_{g,ip})^{-1} \left[\gamma_{\kappa_i} d_{ip} \sum_{n=1}^{N^{\mathrm{CQ}}} x^{\mathrm{CQ}}_{p;\,n} \kappa^{\mathrm{CQ}}_{i;\,n} + +\gamma_g m_g.\right]$$

L'algorithme d'estimation à partir d'un échantillonnage stochastique est modifié en fonction de ce qui est décrit ci-dessus. Pour une itération (k) donnée, les paramètres attachés à une acquisition CQ sont échantillonnés en parallèle jusqu'au niveau « peptide » inclus. Ensuite, on tire un échantillon sous la loi *a posteriori* conditionnelle pour $\mathbf{G}$. L'algorithme s'arrête quand le critère d'arrêt est satisfait. L'Espérance A Posteriori de $\mathbf{G}$, notée $\mathbf{G}_0$, est approchée en moyennant les échantillons tirés sous la loi stationnaire. Le lecteur trouve le pseudo-code de l'algorithme dans Alg. 4.2, page 80.

4.5 Résultats

Nous présentons dans cette section les résultats de la quantification pour les données SRM. Les travaux sur la quantification en mode Full-MS sont résumés dans l'Ann. D, page 151.

4.5.1 Données simulées

Nous évaluons d'abord l'approche d'Inversion-Quantification à l'aide de données simulées. Pour cela, nous considérons une protéine donnant trois peptides, chaque peptide donnant trois transitions. Nous le transcrirons dans la suite du document par l'expression suivante : $\boxed{P = 1} \xrightarrow{1 \quad 3} \boxed{I = 3} \xrightarrow{1 \quad 3} \boxed{L = 9}$. Nous proposons la simulation d'une cohorte à quatre niveaux de concentration différente, les individus étant répartis uniformément sur les classes, imitant des données type « rampe de dilution ». Les distributions ayant servi à établir la cohorte sont $\mathcal{N}(\boldsymbol{x}; m_{\mathsf{m}}, \gamma_{\mathsf{m}})$ avec $(m_1, \gamma_1) = (0.5, 0.025)$, $(m_2, \gamma_2) = (1, 0.05)$, $(m_3, \gamma_3) = (2, 0.1)$, $(m_4, \gamma_4) = (4, 0.2)$. Pour avoir une cohorte suffisamment grande, chaque classe contient 25 individus.

Nous soulignons que l'entrée aurait pu ne pas un mélange de quatre distributions pour imiter une rampe. Pour évaluer les performances, nous aurions pu générer des échantillons de concentration sous une seule distribution d'un support suffisamment grand pour permettre une analyse statistiquement valable. La génération d'une rampe

Alg. 4.1 Résumé de l'algorithme de quantification de protéines

Entrées: $\mathbf{Y}$, paramètres *a priori* pour $\boldsymbol{x}$, $\boldsymbol{\kappa}$, $\boldsymbol{\theta}^{\text{inst}}$
Nombre maximal d'itérations : K^{M}.
Nombre minimal d'itérations de temps de chauffe : K_0.
Nombre d'échantillons pris en compte : K.

1 **fonction** $\left[\hat{\boldsymbol{x}}, \hat{\boldsymbol{\kappa}}, \hat{\boldsymbol{\theta}}^{\text{inst}}\right] = \text{QUANTIFICATION}(\mathbf{Y})$
2 **Initialisation:** $k = 0$, initialiser les paramètres

3 **tant que** $k \leq K^{\text{M}}$ **faire**
4 $k \leftarrow k + 1$
5 échantillonner $(\boldsymbol{\theta}^{\text{inst}})^{(k)} \sim \text{p}(\boldsymbol{\theta}^{\text{inst}} \mid \mathbf{Y}, \boldsymbol{\kappa}^{(k-1)})$
▷ Produit non standard de distributions. Étape de Gibbs hybride.
6 échantillonner $\boldsymbol{\kappa}^{(k)} \sim \text{p}(\boldsymbol{\kappa} \mid \mathbf{Y}, (\boldsymbol{\theta}^{\text{inst}})^{(k)}, \boldsymbol{x}^{(k-1)}$
▷ Distribution normale multivarée. Échantillonnage explicite.
7 échantillonner $\boldsymbol{x}^{(k)} \sim \text{p}(\boldsymbol{x} \mid \boldsymbol{\kappa}^{(k)})$
▷ Distribution normale multivarée. Échantillonnage explicite.
8 **si** Critère d'arrêt du temps de chauffe satisfait **et** $k \geq K_0$ **alors**
9 Poser $K^{\text{M}} \leftarrow k + K$ et $\mathcal{K} := \{k+1, \ldots, k+K\}$
10 **fin si**
11 **fin tant que**

12 **retourner** $\left[\hat{\boldsymbol{x}}, \hat{\boldsymbol{\kappa}}, \hat{\boldsymbol{\theta}}^{\text{inst}}\right] = \dfrac{1}{K} \displaystyle\sum_{k \in \mathcal{K}} \left[\boldsymbol{x}^{(k)}, \boldsymbol{\kappa}^{(k)}, (\boldsymbol{\theta}^{\text{inst}})^{(k)}\right]$
13 **fin fonction**

Alg. 4.2 Résumé de l'algorithme d'étalonnage par échantillon de Contrôle de Qualité.

Entrées: $\boldsymbol{y}_{1:N}^{\mathrm{CQ}}$, concentrations $\boldsymbol{x}_{1:N^{\mathrm{CQ}}}^{CQ}$ connues, paramètres *a priori* for $\mathbf{G}$, $\boldsymbol{\kappa}$, $\boldsymbol{\theta}^{\mathrm{inst}}$
Nombre maximal d'itérations : K^{M}.
Nombre minimal d'itérations de temps de chauffe : K_0.
Nombre d'échantillons pris en compte : K.

fonction $\mathbf{G}_0 = \text{ÉTALONNAGE}\mathrm{CQ}(\boldsymbol{y}_{1:N}^{\mathrm{CQ}})$
  Initialisation: $k = 0$, initialiser les paramètres

  tant que $k \leq K^{\mathrm{M}}$ **faire**
    $k \leftarrow k + 1$
    pour $n = 1, \ldots, N$ **faire**
      échantillonner $(\boldsymbol{\theta}_n^{\mathrm{inst}})^{(k)} \sim \mathrm{p}(\boldsymbol{\theta}^{\mathrm{inst}} \mid \boldsymbol{y}_n, \boldsymbol{\kappa}_n^{(k-1)})$
        ▷ Produit non standard de distributions. Étape de Gibbs hybride.
      échantillonner $\boldsymbol{\kappa}_n^{(k)} \sim \mathrm{p}(\boldsymbol{\kappa} \mid \boldsymbol{y}_n, (\boldsymbol{\theta}_n^{\mathrm{inst}})^{(k)}; \boldsymbol{x}_n^{\mathrm{CQ}})$
        ▷ Distribution normale multivarée. Échantillonnage explicite.
    fin pour

    échantillonner $\mathbf{G}^{(k)} \sim \mathrm{p}(\mathbf{G} \mid \boldsymbol{\kappa}_{1:N^{\mathrm{CQ}}};\ \boldsymbol{x}_{1:N^{\mathrm{CQ}}}^{\mathrm{CQ}})$
      ▷ Distribution normale multivariée. Échantillonnage explicite.
    si Critère d'arrêt du temps de chauffe satisfait **et** $k \geq K_0$ **alors**
      Poser $K^{\mathrm{M}} \leftarrow k + K$ et $\mathcal{K} := \{k+1, \ldots, k+K\}$
    fin si
  fin tant que

  retourner $\mathbf{G}_0 = \dfrac{1}{K} \sum_{k \in \mathcal{K}} \mathbf{G}^{(k)}$
fin fonction

permet cependant premièrement d'être proche des campagnes expérimentales réelles en protéomique quantitative, et deuxièmement d'analyser le comportement ponctuel du quantifieur.

Quant aux paramètres de simulation de données, le gain de préparation ψ vaut 1, la matrice de digestion est bloc-unitaire, la digestion a un rendement de 1. Les autres paramètres instrumentaux ont été choisis de manière à imiter la variabilité technologique. Les positions de pic chromatographique sont distribuées uniformément dans l'intervalle de l'observation, les largeurs associées uniformément entre 20 et 100 min^{-2}. Par souci de simplicité, le gain de transition est unitaire pour tous les fragments ; le gain global cependant varie selon un distribution normale centrée en 1000 et de précision $1/50^2$, équivalent à un écart type de 50, c'est-à-dire une fluctuation de 10 % environ au sens plus/moins deux écarts types. Chaque transition est perturbé par un bruit blanc centré de précision $5 \cdot 10^{-5}$, résultant en un rapport signal sur bruit moyen de 2.3 dB.

Pour obtenir les estimations et au vu de la complexité des calculs, nous choisissons un nombre d'itérations $K = 5000$, nous nous servons des dernières 2000 itérations comme échantillons de la loi cible pour le calcul de l'estimateur. L'estimation est retournée après 6 s de calcul. Les lois *a priori* aux extrémités de la hiérarchie sont toutes vaguement informatives, laissant ainsi les données « s'exprimer ».

Présentation

La Fig. 4.3 illustre les résultats de la performance de l'Inversion-Quantification sur des données linéarité simulées. Nous y traçons sur l'abscisse la concentration vraie $x^\star$ avec laquelle la donnée est générée, et sur l'ordonnée la concentration estimée $\hat{x} = \psi^{\mathsf{q}}(\mathbf{Y})$. Chaque point $(x_n^\star, \hat{x}_n)$ correspond au couple vérité/estimation de l'individu n. L'intervalle de 5%-crédibilité a été ajouté par l'adjonction de barres à deux fois l'écart type empirique de la distribution marginale *a posteriori* de la concentration. La droite de régression est dessinée en vert, la bissectrice en noir et pointillé.

Globalement, la figure illustre une bonne performance de la méthode. La droite de régression, obtenue en minimisant le critère quadratique $\mathbb{J}$ imposé pour la régression simple, se confond quasiment avec la bissectrice, *i.e.* le modèle de prédiction est l'identité. De plus, le coefficient de détermination R^2 a une valeur de 0.99 qui est indéniablement proche de 1. Dans les conditions des simulations, le quantifieur est à peu près un estimateur parfait (voir Déf. ⟨2.26⟩, page 30) qui, d'après la Proposition ⟨2.29⟩ atteint une droite de régression se confondant avec la bissectrice avec $R^2 = 1$.

Nous venons d'analyser la droite de régression par rapport aux valeurs de concentration connues. Or, nous avons simulé une rampe de dilution de laquelle nous connaissons uniquement la valeur nominale par dilution, les valeurs de concentration vraies nous étant inconnues, comme nous le verrons plus loin. La droite de régression d'une telle étude, confrontant la référence de dilution à l'estimation, est donnée par l'expression linéaire $\hat{x} = 0.014 + 0.99x$ qui atteint un coefficient de détermination de $R^2 = 0.98$.

Le coefficients de variation, synonyme de dispersion normalisée, sont faibles par classe : 0.23, 0.11, 0.083 et 0.062 respectivement. Les biais s'annulent, les erreurs quadratiques sont peu élevées. Toutes les mesures d'erreur, prises par rapport à la moyenne m_{m} de chaque classe, sont résumées dans la Tab. 4.1. Nous pouvons ainsi quantifier l'impact de la variabilité biologique que nous avons créée en simulant des distributions de concentration non réduites à des distributions de Diracs.

Les barres d'erreur qui indiquent l'intervalle de 5%-crédibilité pour chaque estimation, obtenue à partir des échantillons de la loi marginale *a posteriori* pour x, n'ont que pour des cas isolés aucune intersection avec la bissectrice ; autrement dit, les valeurs vraies sont contenues dans les intervalles de 5%-crédibilité de presque toutes les esti-

Tab. 4.1 Tableau des mesures d'erreur pour l'évaluation de performance, utilisant des données simulées.

droite de régression par rapport à $x^\star$	α	-0.00055
	β	1.0053
	R^2	0.99168
droite de régression par rapport à m_{m}	α	0.0140
	β	0.9900
	R^2	0.98446
coefficients de variation	$\mathbb{CV}_{\boldsymbol{y}\|x,m_1}$	0.232
	$\mathbb{CV}_{\boldsymbol{y}\|x,m_2}$	0.111
	$\mathbb{CV}_{\boldsymbol{y}\|x,m_3}$	0.0833
	$\mathbb{CV}_{\boldsymbol{y}\|x,m_4}$	0.0615
biais	$\mathbb{B}_{\boldsymbol{y}\|x,m_1}$	-0.0106
	$\mathbb{B}_{\boldsymbol{y}\|x,m_2}$	-0.00143
	$\mathbb{B}_{\boldsymbol{y}\|x,m_3}$	0.0364
	$\mathbb{B}_{\boldsymbol{y}\|x,m_4}$	-0.0433
biais moyen	$\mathbb{B}_{x,\boldsymbol{y}}$	-0.00475
variance	$\mathbb{V}_{\boldsymbol{y}\|x,m_1}$	0.0124
	$\mathbb{V}_{\boldsymbol{y}\|x,m_2}$	0.0118
	$\mathbb{V}_{\boldsymbol{y}\|x,m_3}$	0.0276
	$\mathbb{V}_{\boldsymbol{y}\|x,m_4}$	0.0568
variance moyenne	$\mathbb{V}_{x,\boldsymbol{y}}$	0.0272
erreur quadratique	$\mathbb{EQ}_{\boldsymbol{y}\|x,m_1}$	0.0125
	$\mathbb{EQ}_{\boldsymbol{y}\|x,m_2}$	0.0118
	$\mathbb{EQ}_{\boldsymbol{y}\|x,m_3}$	0.0290
	$\mathbb{EQ}_{\boldsymbol{y}\|x,m_4}$	0.0587
erreur quadratique moyenne	$\mathbb{EQ}_{x,\boldsymbol{y}}$	0.0272

mations malgré la présence d'un fort bruit de mesure.

Discussion

Les résultats de la Fig. 4.3 complétée par la Tab. 4.1 montrent de très bonnes performances malgré une puissance de bruit élevée : les erreurs sont petites par rapport à la variabilité injectée, les intervalles de crédibilité contiennent les valeurs vraies et l'estimateur est proche de l'estimateur parfait. Sur ce jeux de données, l'approche Inversion-Quantification est validée.

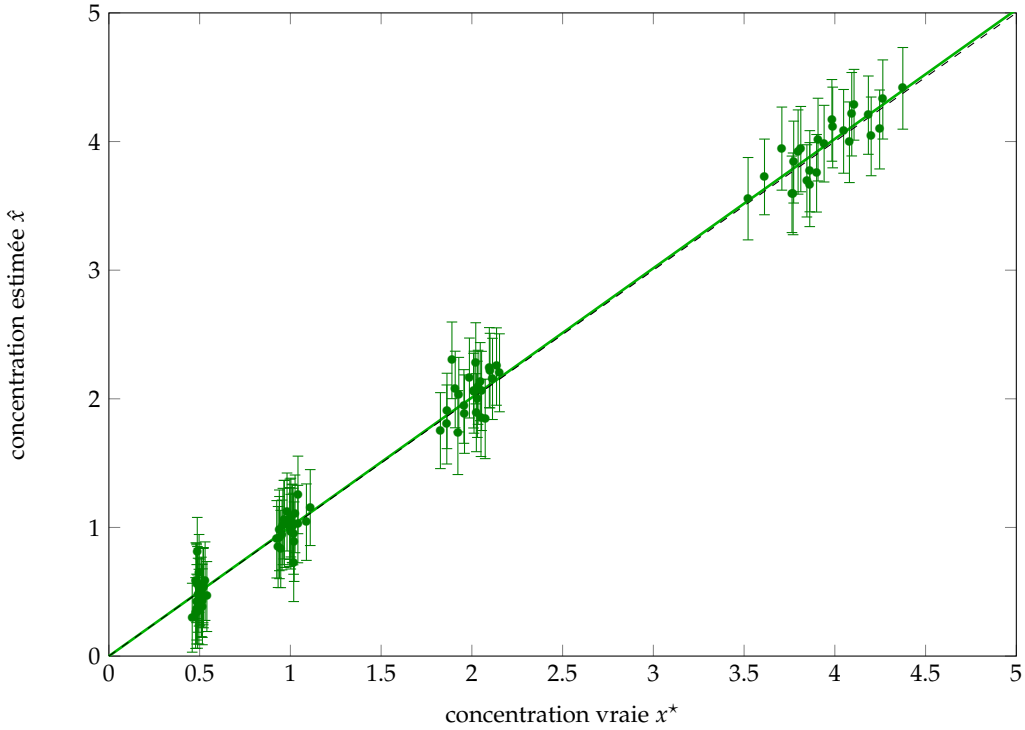
Cependant, gardons à l'esprit que nous pouvons avoir un effet de surmodélisation (*overfitting*). En effet, on commet un « crime inverse » : la génération des données est faite avec exactement le même modèle que celui utilisé pour l'inversion. Or, ce modèle n'est qu'approximatif, surtout en ce qui concerne le bruit de mesure — modélisé blanc gaussien, en réalité non négatif et possiblement corrélé. Même si les résultats peuvent sous cet angle avoir l'air flatteur, nous verrons dans la suite en analysant les données réelles que la surmodélisation est négligeable.

4.5.2 Données synthétiques

Nous proposons d'évaluer dans cette section les performances de l'Inversion-Quantification par rapport au jeu de données synthétiques, voir Sect. 3.6. Pour l'analyse, nous choisissons un sous-protéome de huit protéines des quarante incluses dans les données : Ezrine, HSP60, HSP71, IFABP, LFABP, PDI, Villine, Protéine X[4].

4. NB — Le nom *Protéine X* a été sélectionné pour des raisons de confidentialité.

Fig. 4.3 Présentation de la régression vérité ÷ estimation sur des données SRM simulées



Pour chaque échantillon CQ, nous disposons de deux mesures, la troisième étant trop corrompue et donc éliminée, comme expliqué précédemment dans Sect. 3.6. Il est ainsi difficile de tirer des conclusions statistiquement valables à cause du faible nombre de mesures. C'est la raison pour laquelle nous omettons la mention du coefficient de détermination. Cependant, nous nous en contentons pour voir les « tendances » du quantifieur.

Le critère d'arrêt de l'algorithme de quantification est un nombre d'itérations maximales $K = 5000$ atteint, les dernières deux mille servant d'échantillons de la loi cible. Le temps de calcul moyen pour un échantillon de cette cohorte à huit protéines est chiffré à 73 s.

Présentation

L'évaluation est illustrée à l'aide de droites de régression qui sont données dans les Figs. 4.4 et 4.5, chacune des huit est présentée sur une sous-figure. Globalement, on peut constater une bonne linéarité. Les exceptions, *i.e.* les points éloignés de chaque droite de régression, se trouvent dans l'échantillon CQ4 à la fois dans l'étude de l'Ezrine, de l'HSP60 et de la Villine où on remarque une sous-expression. La dispersion de ces points étant extrêmement faible, nous pouvons relier cette sous-expression à un aléa de la préparation de la dilution.

Ensuite, les droites de régression ont une pente β comprise entre 0.8 et 1.2 pour les protéines Ezrine, HSP60, IFABP, LFABP, Villine et Protéine X ; un peu plus éloignées sont les pentes des protéines HSP71 (0.74) et PDI (0.78). La première des deux montre aussi une concentration endogène de 0.19 alors que les autres protéines ont un niveau endogène négligeable (< 0.1). Chez la protéine PDI, on constate également qu'à partir du CQ3, un des deux points de mesures semble sous-exprimé (CQ3 : 0.254 vs. 0.176 ; CQ4 : 0.494 vs. 0.336, CQ5 : 1.001 vs 0.652), l'autre étant estimé à une valeur correcte par rapport au niveau de dilution.

Individuellement, chaque estimation est accompagnée de son intervalle de 5%-crédibilité. La figure illustre également les incertitudes sur les estimations selon l'écart type empirique. Pour de petites concentrations l'incertitude est très petite ; plus la concentration protéique croît, plus l'incertitude est grande.

La Fig. 4.6 présente les droites de régression dans un repère double-logarithmique pour la protéine LFABP (en tant que représentante des autres protéines où la linéarité est confirmée) et la Villine où nous avons remarqué une sous-expression. Pour cette étude, nous avons dû éliminer les échantillons CQ0, affichant un rapport de 0 dont le logarithme n'est pas défini. Compte tenu du faible nombre d'acquisitions traitées, on voit que le coefficient de la log-log-régression de la LFABP est proche de 1, démontrant le caractère linéaire de la proportionnalité entre rapports de dilution injecté et estimé. Pour la Villine, on trouve une droite de log-log-régression de $\log_2(y) = -0.06 + 1.50 \log_2(x)$, *i.e.* la dépendance entre entrée et sortie ne se décrit pas d'une manière linéaire, mais en fonction de la puissance 1.5 de l'entrée. La courbe de régression résultante dans un repère cartésien est illustrée dans la Fig. 4.7.

Enfin, la Fig. 4.8 présente un exemple de reconstruction de données. Nous superposons les acquisitions et les sorties modèles calculées à partir des estimations des paramètres. Chaque ligne regroupe les trois traces d'un même peptide de la protéine LFABP. On constate généralement une bonne concordance entre les signaux, la reconstruction approchant bien le signal obtenu. Les paramètres chromatographiques sont bien estimés. L'utilisation de la redondance des informations a permis d'écarter les pics parasites dans l'estimation dans les traces associées au premier et au deuxième peptide. La qualité des données du troisième peptide est très bonne, les perturbations sont presque inexistantes. L'estimation associée est – à l'image des traces – également très bonne.

Discussion

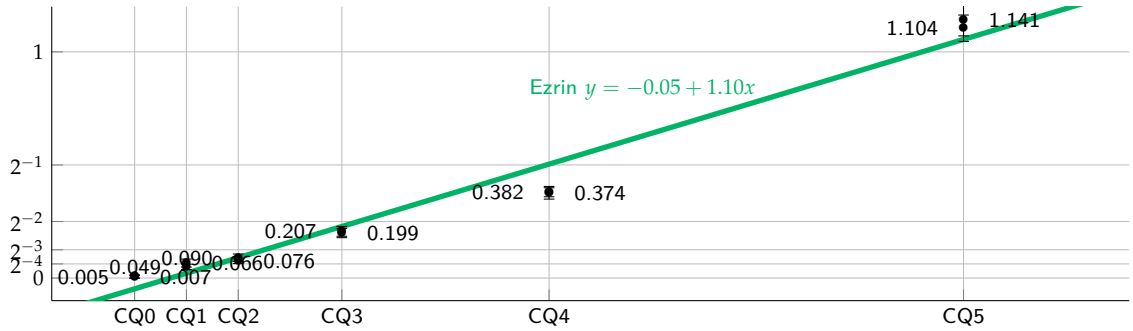
L'illustration des résultats dans les Figs. 4.4 et 4.5 valide — au moins de manière « tendancielle » — la méthode d'Inversion-Quantification pour le mode SRM à partir de données expérimentales de type « rampe de dilution ». En effet, on retrouve une bonne linéarité avec peu de points non concordants. Ceux-ci apparaissent soit dans un couple d'échantillons de dilution précis (CQ4), soit pour un échantillon du couple (PDI, CQ3 à CQ5).

La protéine HSP71 semble sortir un peu de cette vue générale. La pente de sa droite de régression vaut 0.74, l'ordonnée à l'origine vaut 0.19. À part le CQ5, les autres échantillons de cette protéine sont sur-exprimés par rapport à la dilution attendue. Comme cet effet n'apparaît nulle part ailleurs, on peut se poser des questions quant à la préparation et l'injection de cette protéine, sachant que les estimations régressent bien linéairement.

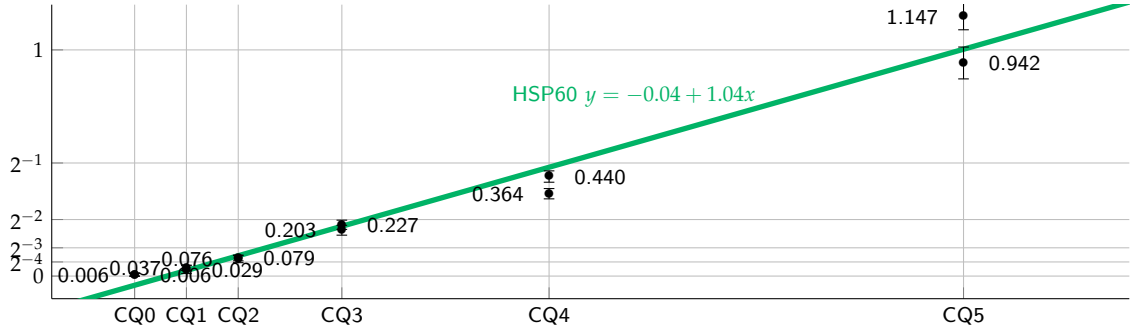
Dans la Fig. 4.7, nous avons ajouté à la régression linéaire la courbe de régression de puissance 1.5 issue de l'analyse en échelle double-logarithmique. Nous observons que cette régression n'est pas plus satisfaisante que la régression linéaire. On souligne à nouveau que le nombre de mesures est faible et qu'il nous faudra un nombre plus élevé pour tirer des conclusions valables dans un sens statistique.

Quant à l'objection de sur-modélisation que nous avons évoquée précédemment, on voit dans les résultats sur ces données réelles que le fait de traduire le bruit de mesure par un bruit blanc gaussien centré semble ne pas affecter (ou affecter négligemment) les estimations, selon ce qu'on peut « conclure » avec un petit nombre de données, confirmé également par les bonnes reconstructions, Fig. 4.8.

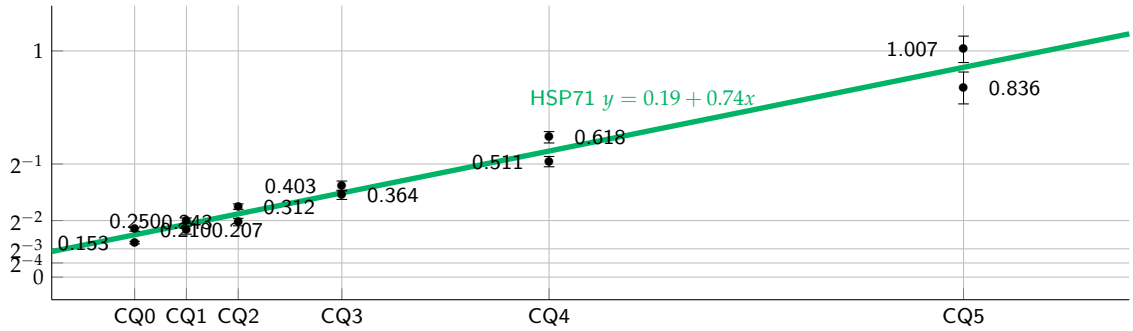
Fig. 4.4 Droites de régression pour les quatre premières protéines du jeu de données « linéarité » : (a) Ezrine, (b) HSP60, (c) HSP71, (d) IFABP. Sur l'abscisse les « quantités » vraies (en termes de CQ), sur l'ordonnée les rapports de dilution estimés. *(Suite dans la Fig. 4.5)*



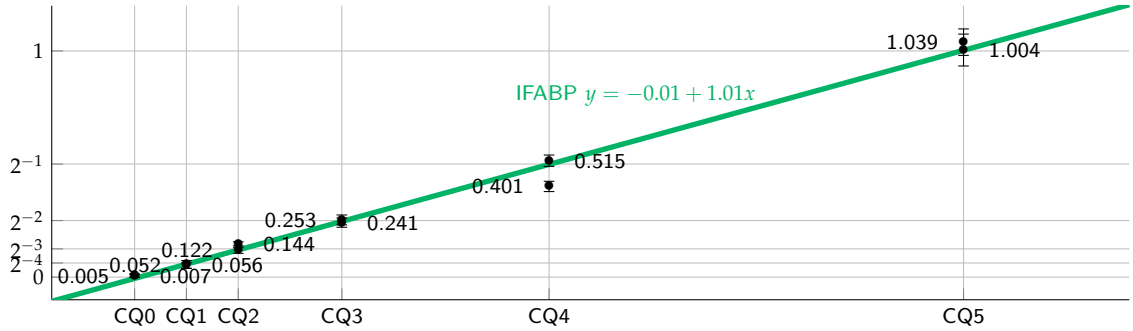
(a) Ezrine



(b) HSP60

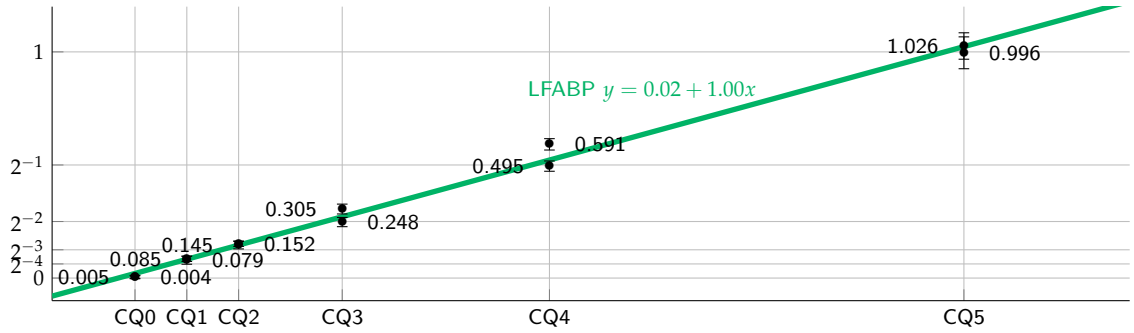


(c) HSP71

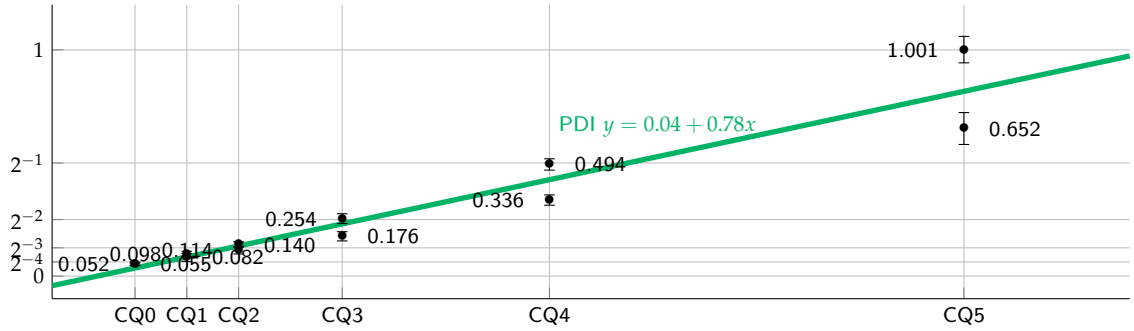


(d) IFABP

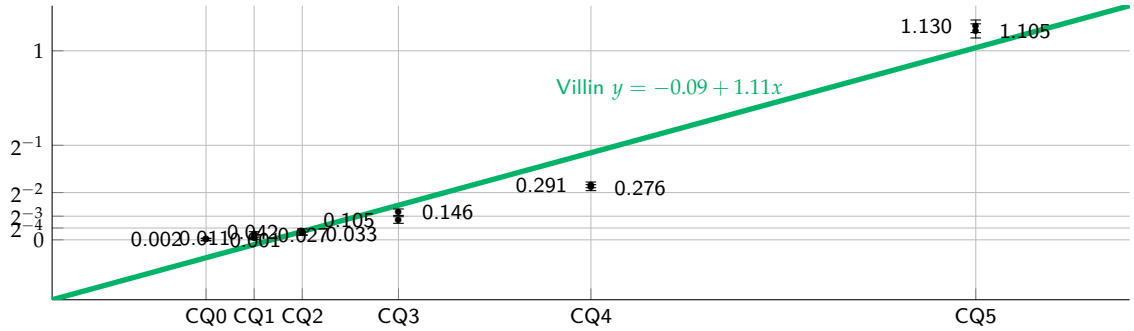
Fig. 4.5 *Suite de la Fig. 4.4* : Droites de régression pour les quatre dernières protéines du jeu de données « linéarité » : (a) LFABP, (b) PDI, (c) Villine, (d) Protéine X. Sur l'abscisse les « quantités » vraies (en termes de CQ), sur l'ordonnée les rapports de dilution estimés.



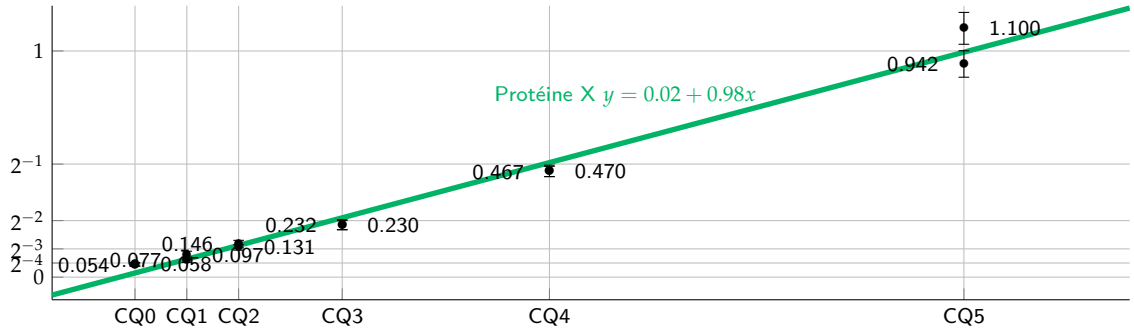
(a) LFABP



(b) PDI

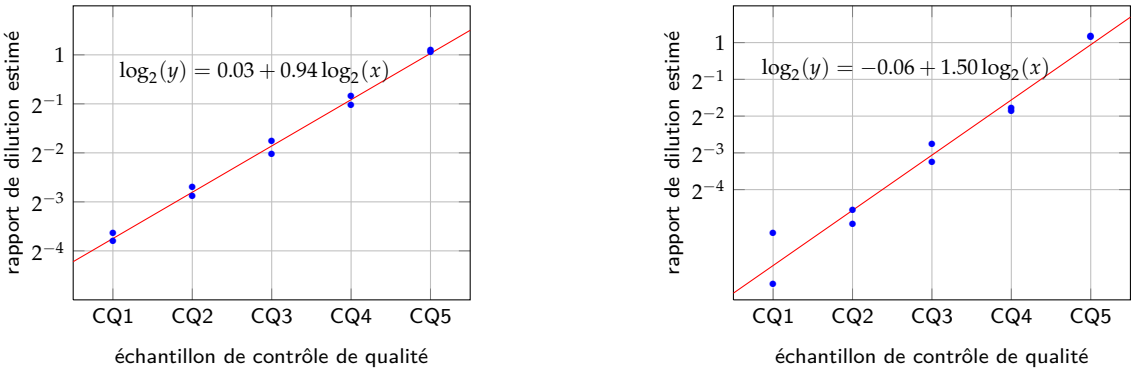


(c) Villine



(d) Protéine X

Fig. 4.6 Représentation des régressions sur les données synthétiques en échelle double-logarithmique pour les protéines LFABP et Villine



(a) LFABP ($\log_2/\log_2$) (b) Villine ($\log_2/\log_2$)

Fig. 4.7 Représentation des régressions linéaire (vert) et exponentielle (rouge) de la Villine

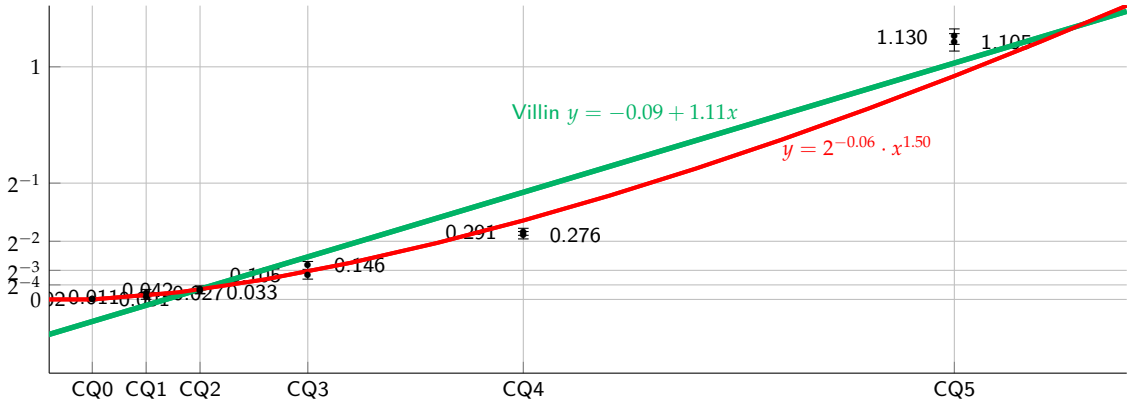
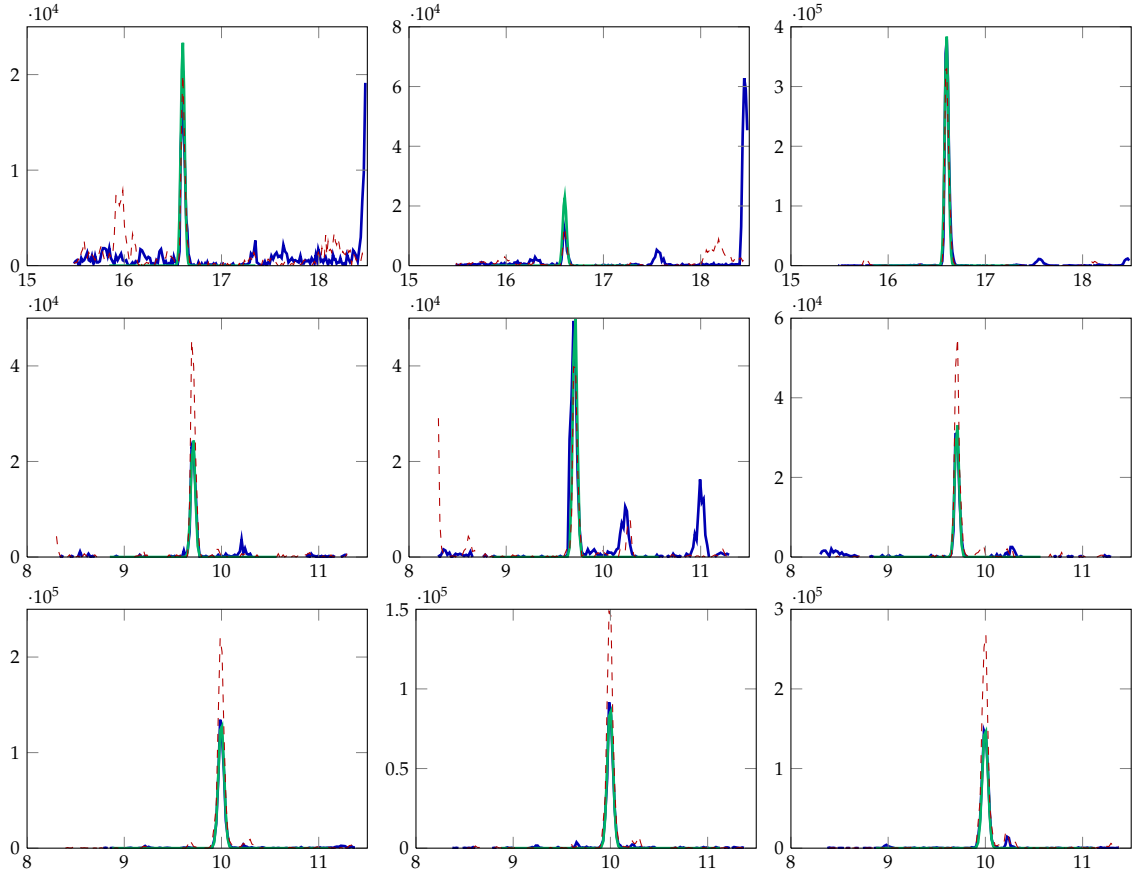


Fig. 4.8 Reconstruction d'une donnée SRM du jeu de données synthétiques pour la protéine LFABP. (— signal enregistré, — signal reconstruit, – – signal du standard.) Chaque ligne correspond à un même peptide, chaque figure à une trace associée au peptide. L'abscisse correspond au temps de rétention en minutes, l'ordonnée à l'intensité du signal.



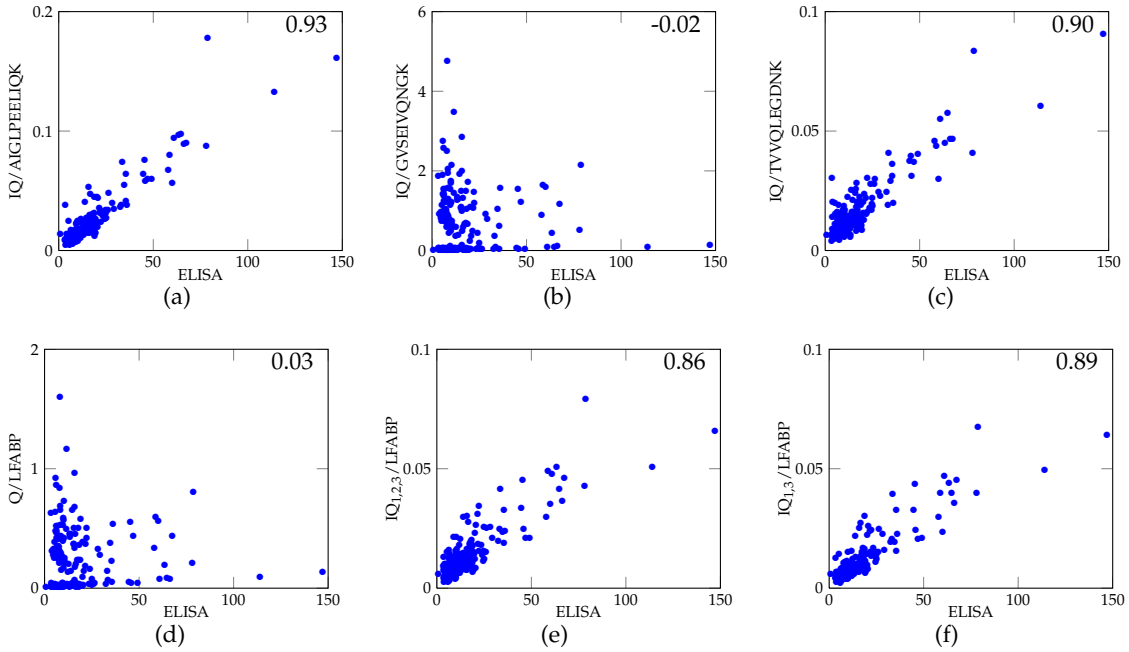
4.5.3 Données cliniques

Les données à disposition pour l'évaluation de notre méthode comportent également des données cliniques issues d'une campagne expérimentale sur des patients du cancer colorectal et sur des individus contrôle. La difficulté ici vient clairement du fait qu'on ne peut avoir de concentration vraie. Nous pouvons donc seulement comparer à d'autres estimations. Nous avons discuté ce problème dans Sect. 3.6 et motivé l'utilisation des tests ELISA comme référence. Malheureusement, une seule protéine est dosée à la fois en ELISA et en SRM. En prenant la première comme référence, supposée sans erreur, nous analysons la corrélation entre les estimations ELISA et SRM, les estimations n'étant pas à la même échelle pour une analyse par droite de régression.

Les données sont au préalable étalonnées par les échantillons de Contrôle de Qualité. Si la date d'acquisition de données cliniques correspond à une date pour laquelle nous disposons d'échantillons CQ, c'est eux que nous utilisons pour l'étalonnage ; sinon, nous calculons une moyenne sur les CQ de tous les jours disponibles.

Le temps de calcul pour l'obtention des estimations est de 20 s par donnée en considérant $K = 5000$ itérations incluant un temps de chauffe de $K_0 = 3000$. L'algorithme s'arrête en atteignant le nombre d'itérations K.

Fig. 4.9 Corrélations entre les tests ELISA et les estimations de l'Inversion-Quantification : (a) pour le peptide AIGLPEELIQK, (b) pour le peptide GVSEIVQNGK, (c) pour le peptide TVVQLEGDNK, (d) pour la protéine en utilisant une moyenne arithmétique des estimations peptidiques, (e) pour la protéine tenant compte des trois peptides ($IQ_{1,2,3}$), (f) pour la protéine en excluant le peptide corrompu de l'estimation ($IQ_{1,3}$). En haut à droite de chaque figure se trouve la valeur de la corrélation.



Présentation

La Fig. 4.9 présente graphiquement les corrélations entre les estimations ELISA, sur l'abscisse de chaque sous-figure, et les estimations issues de l'Inversion-Quantification (IQ) que nous avons exposée dans ce chapitre.

Par la méthodologie choisie, nous avons accès non seulement à la variable d'intérêt (concentration protéique), mais aussi aux paramètres de nuisance qui peuvent être analysés. Ainsi, nous considérons dans les sous-figures 4.9(a), (b) et (c) les corrélations avec les estimations au niveau peptidique. On voit que les estimations IQ des peptides AIGLPEELIQK et TVVQLEGDNK corrèlent très bien avec les estimations ELISA, les corrélations valant 0.93 et 0.90 respectivement.

Le peptide GVSEIVQNGK par contre montre une corrélation quasi nulle. En effet, la plupart des valeurs estimées sont beaucoup plus grandes et sur une autre échelle que celles des autres peptides alors que les peptides sont censés mesurer la même quantité (modulo l'étalonnage). On peut certes deviner un groupe de points qui semblent être corrélés avec les tests ELISA vers les valeurs faibles de l'ordonnée, mais la plupart des points sont bien au dessus de ce niveau. Nous y reviendrons dans la discussion des résultats.

Dans la sous-figure 4.9(d), nous présentons la corrélation entre le test ELISA et la quantification protéique utilisant la moyenne arithmétique des estimations peptidiques. On voit clairement que, même si on distingue une certaine linéarité des points en bas de la figure, la plupart des estimations sont perturbées par le résultat du deuxième peptide. La corrélation est chiffrée à 0.03, suggérant qu'il n'y a presque pas de relation linéaire entre les valeurs.

La sous-figure 4.9(e) illustre ensuite la corrélation des estimations ELISA et la recons-

truction protéique IQ sans tenir compte de l'inadaptation du peptide GVSEIVQNGK. On atteint une corrélation de 0.86 ce qui est toujours un bon résultat. Comme l'estimation protéique par l'IQ n'est pas une simple moyenne arithmétique mais prend en compte la qualité des données et de précision de la distribution *a posteriori* conditionnelle sur les peptides, l'impact du deuxième peptide est atténué. La très mauvaise corrélation du peptide GVSEIVQNGK dégrade seulement de manière faible la concentration protéique par rapport aux corrélations des deux autres peptides. Le « poids » attribué par la méthode aux peptides peut être mesurée en considérant les corrélations inter-estimations : alors que les coefficients de corrélation entre l'estimation de la concentration protéique et des quantités des peptides AIGLPEELIQK et TVVQLEGDNK sont de 0.942 et 0.887 respectivement, il n'est que de 0.0756 pour le peptide GVSEIVQNGK. Nous mettons en relief que l'attribution du « poids » ne se fait pas en considérant une étude statistique sur une cohorte de mesures ; la méthode n'a pas d'autres entrées que la donnée à quantifier (et éventuellement des paramètres supplémentaires). Ainsi, l'algorithme montre un caractère autocalibrant en détectant par lui-même l'inadéquation d'une mesure peptidique. Ajoutant cependant que l'élimination manuelle du peptide dans l'estimation mène naturellement à une meilleure corrélation (0.89) comme nous montrons dans la Fig. 4.9(f).

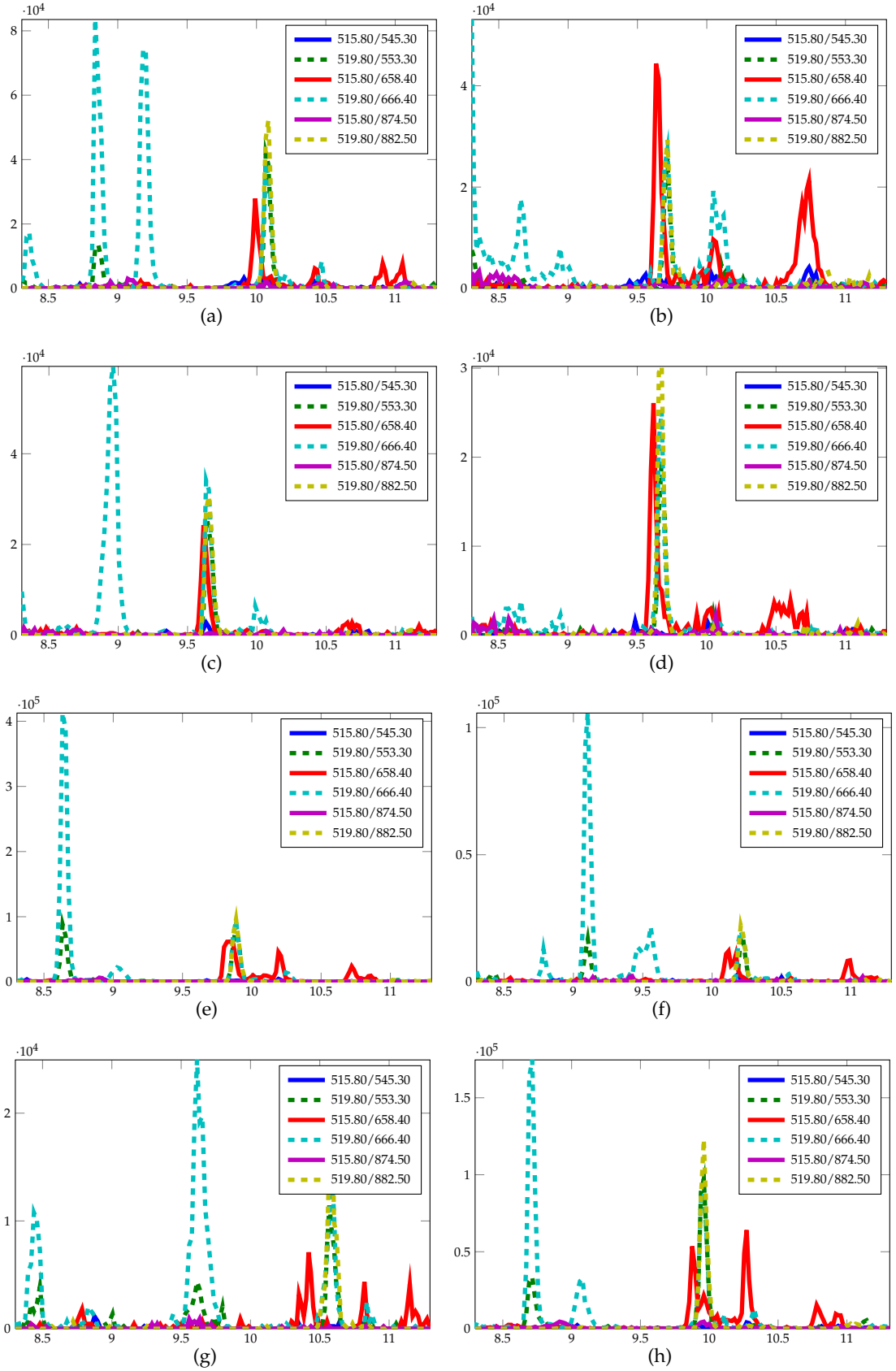
Discussion

Nous avons remarqué dans la description des résultats qu'un des peptides, GVSEIVQNGK, présente un comportement anormal par rapport aux autres peptides et aux attentes. Pour comprendre d'où peut venir ce défaut, on peut accuser les développements théorique et algorithmique ou les observations. L'algorithme ayant donné des résultats convaincants sur la même protéine en utilisant des données synthétiques, nous proposons de regarder les données. Dans la Fig. 4.10, nous présentons les traces du peptide GVSEIVQNGK pour une sélection de huit individus. On voit que les données sont perturbées par des contaminants qui affectent surtout la transition 515.80/658.40 (rouge) sous des formes et positions différentes. Ces « parasites » ne sont pas expliqués par le modèle mis en place, notamment en ce qui concerne le caractère commun des pics chromatographiques. Il se trouve également qu'il est moins coûteux en termes d'erreur et, donc, de vraisemblance, d'adapter la forme du pic perturbé et de négliger les deux pics non perturbés.

L'exclusion du peptide GVSEIVQNGK améliore certes la corrélation (augmentation de 0.03 points), mais elle reste *inférieure* à celles des peptides pris individuellement. On s'attendrait à ce que la combinaison des valeurs peptidiques apporte un lissage des fluctuations et que la corrélation résultante soit supérieure ou égale à celle du meilleur peptide. Or, ce n'est pas le cas ici. Les peptides étant marqué par un standard lourd, les protéines sont ensuite inférées par un étalonnage CQ préalable. En faisant ainsi, nous rapportons un gain appris à l'aide des données extérieures sur le traitement numérique en cours. Même si ce gain est supposé stable, cette analyse de résultats montre que l'étalonnage tel qu'il est conçu n'est pas suffisant et nécessite des améliorations, actuellement en phase de réflexion.

Malgré ceci, nous pouvons valider l'approche Inversion-Quantification sur données réelles en s'appuyant sur le fait qu'une corrélation de 0.86 (voire 0.89) est satisfaisante, surtout en comparant les temps d'acquisitions et d'estimation des méthodes. Avec ces valeurs, nous sommes proche de la corrélation d'environ 0.9 considérée comme validant deux tests ELISA entre eux.

Fig. 4.10 Présentation des traces GVSEIVQNGK pour une sélection de huit individus. Sur l'abscisse, temps de rétention en minutes ; sur l'ordonnée, l'amplitude du signal.



4.6 Conclusion

Dans ce chapitre, nous avons présenté les développements théorique et algorithmique de l'Inversion-Quantification, son but étant de quantifier les protéines d'un échantillon biologique à partir d'une donnée chromatographique/spectrométrique. Pour cela, nous avons construit la loi jointe et exprimé l'estimateur (« quantifieur ») dans le cas présent. Afin de mettre en œuvre un algorithme d'échantillonnage stochastique type MCMC adoptant une structure de Gibbs, nous avons calculé les lois *a posteriori* conditionnelles pour chacun des paramètres pour les deux modes de spectrométrie que nous considérons.

Le standard utilisé dans les données SRM étant l'AQUA, le lien entre protéines et peptides ne peut être fait de manière interne. Nous avons exposé l'étalonnage par échantillons de Contrôle de Qualité, incluant les aspects conception et développement algorithmique.

Finalement, nous avons confronté la méthode à des données diverses pour évaluer l'Inversion-Quantification et pour la valider. Nous avons considéré pour cela des données simulées selon notre modèle direct, des données réelles de type rampe de dilution affichant une linéarité, et des données réelles cliniques en comparant les estimations avec une référence. Nous avons validé l'approche dans l'intégralité des expériences, avec des réserves quant à la taille de la cohorte pour les données synthétiques. L'analyse des données cliniques a appelé deux remarques. Premièrement, nous avons constaté qu'un peptide se comporte d'une manière inattendue. La considération des traces associées à ce peptide a montré qu'elles sont fortement corrompues. Ensuite, nous avons remarqué une dégradation de la corrélation au niveau de l'estimation de la concentration protéique, probablement suite à l'utilisation de l'étalonnage CQ dans les données cliniques. Nous avons également découvert une forte perturbation des données du peptide GVSEIVQNGK de la protéine LFABP, impactant les résultats d'estimations mais de manière très modérée. Parmi les traces du peptide concerné, une seule était affectée, et l'élimination manuelle du peptide corrompu associé a amélioré les résultats. L'amélioration de l'étalonnage par d'autres moyens que la simple utilisation des échantillons de Contrôle de Qualité est actuellement en phase de discussion à l'intérieur de l'équipe et du consortium BHI-PRO.

5 Inversion-Classification

Dans ce chapitre, nous allons voir l'apport majeur de cette thèse : une méthode de classification issue de la méthodologie des problèmes inverses. Elle ne se limite pas aux instruments LC-MS ou SRM, mais est extensible presque à d'autres problèmes de nature hiérarchique. C'est la raison pour laquelle nous n'allons pas raisonner en paramètres associés à tel ou tel instrument, mais par niveau hiérarchique, à l'instar du chapitre précédent. On utilisera à nouveau les vocables « paramètre d'intérêt » et « paramètres de nuisance ».

Pourquoi coupler « classification » et « problèmes inverses » ? La méthodologie des problèmes inverses permet une inférence *jointe* de *tous* les paramètres en jeu, tant les paramètres continus (les paramètres attachés aux instruments, la concentration des protéines) que les paramètres discrets (la classe). Elle permet également de rendre plus robuste l'estimation grâce à l'information apportée par la modélisation du problème direct. Le cadre bayésien nous sera ensuite utile pour séparer et quantifier les sources d'incertitude.

5.1 Classification : apprendre et classer

Le terme *classification* est souvent confondu avec l'action de *classer une nouvelle observation*. Mais il y a plus que cela. Attaché à la classification, il y a aussi l'action d'*apprendre les paramètres caractéristiques* des classes afin de pouvoir prendre une décision, afin d'estimer la classe d'appartenance. Ces deux actions reposent sur des choix qui sont faits : quels paramètres caractéristiques, quelles fonctions d'estimation, quelle cohorte d'apprentissage et à quel effectif, ... Les réponses à ces questions structurent le *classifieur*, *i.e.* la « routine » qui propose l'estimation de la classe d'une nouvelle observation, compte tenu de l'apprentissage et d'autres informations. La classification est donc composée de deux tâches :

Apprendre les caractéristiques — Premièrement, il est nécessaire d'entraîner le classifieur. Pour cela, on utilise généralement des cohortes d'individus de classe déterminée sur lesquelles on apprend les caractéristiques (*features*) qui sont censées séparer les classes. On appellera ces caractéristiques aussi la *variable explicative pour la classification*. Cette étape *entraîne* le classifieur.

Estimer la classe — Deuxièmement, à l'aide du classifieur entraîné, on peut associer à un nouvel individu de classe indéterminée l'estimation de sa classe. On dit aussi qu'on *applique* le classifieur à une nouvelle observation.

Considérons le *premier choix* que nous faisons pour structurer le classifieur : adresser les deux tâches séparément. Pour quelles raisons pourrait-on avoir envie de procéder comme ça ? *Primo*, nous pouvons avoir des cohortes dont on veut seulement connaître les caractéristiques, *e.g.* la distribution, pour des buts de comparaison. *Secundo*, les caractéristiques ainsi apprises sont réutilisables dans des cas où on veut estimer la classe de plusieurs nouvelles observations à des temps différents. Un « ré-apprentissage » n'est pas nécessaire. *Tertio*, mettons nous dans les conditions suivantes : les caractéristiques correspondent aux hyperparamètres de la distribution de la concentration protéique. Au temps t, nous disposons d'une première cohorte d'apprentissage C_1 sur laquelle nous entraînons le classifieur. Au temps $t + \Delta t$ nous disposons de la fusion F de la cohorte

précédente avec une nouvelle cohorte C_2. Apprendre les caractéristiques à partir de F est équivalent à apprendre les caractéristiques à partir de C_2 où les paramètres *a priori* correspondent aux caractéristiques apprises auparavant sur la cohorte C_1. *Quarto*, étant donné un grand nombre d'individus par cohorte, l'impact de la variabilité dans l'apprentissage semble moins important.

Notons cependant que la méthodologie bayésienne (comme d'autres) permet également de considérer un problème joint « apprendre *et* classer ». Pour cela, on déduit la probabilité *a posteriori* pour la classe sachant l'observation à classer et toutes les observations d'apprentissage. Ce faisant, nous marginalisons tous les paramètres de nuisances, *i.e.* paramètres instruments et concentrations protéiques associés à chaque observation. Par la marginalisation, nous intégrons toutes les variabilités, y compris celles issues de l'inversion des données d'apprentissage, aussi faibles soient-elles.

En étendant ces travaux, cette stratégie est non seulement valable pour une simple classification, mais aussi pour un étiquetage semi-automatique, c'est-à-dire un apprentissage semi-supervisé avec des étiquettes manquantes : imaginons que nous avons à disposition une cohorte à étiquette déterminée sans erreur et une cohorte à étiquette indéterminée. À l'aide d'une modélisation adéquate, on peut conjointement attribuer une classe aux individus de la cohorte à étiquette indéterminée et estimer les paramètres de la distribution des concentrations, mettant à jour ainsi le partitionnement de la cohorte.

Le *deuxième choix* quant à la structure du classifieur consiste à définir les caractéristiques. Elles déterminent l'espace dans lequel le classifieur agit et travaille.

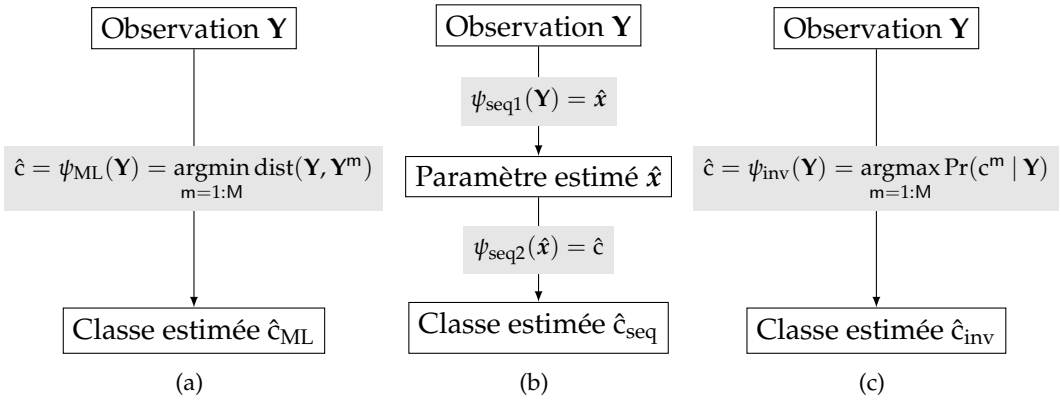
Machine Learning Dans les approches *Machine Learning*, la classification est entreprise dans l'*espace des données ou des caractéristiques* (feature space), *cf.* Fig. 5.1(a). Ceci ne prend en compte aucun modèle physique explicite : l'observation est simplement comparée à des données (*i.e.* observations ou caractéristiques) représentatives de chaque classe. Le classement est ensuite fait à partir de la minimisation d'un coût. Ainsi, c'est sur la conséquence finale d'un événement que l'on travaille et c'est elle que l'on utilise pour classer.

Dans les Interfaces Cerveau/Machine (ICM), Barachant et al. [80] classent les intentions d'un utilisateur, connecté à une machine *via* un dispositif de lecture d'activité électrique du cerveau. Les signaux obtenus sont modélisés par des processus aléatoires, gaussiens, centrés et de covariance différentes selon l'intention. Ensuite, les matrices de covariance « vivant » dans la géométrie riemannienne, les auteurs estiment le centroïde de chaque classe à l'intérieur de cette géométrie. La classification se fait ensuite en minimisant la distance riemannienne de l'échantillon à classer par rapport aux centroïdes représentatifs des classes.

Approche Séquentielle Si un modèle d'acquisition peut être mis en place, on peut procéder à des estimations intermédiaires avant de classer à partir de ces dernières, réalisant ainsi une *approche séquentielle* comme présenté dans la Fig. 5.1(b). On applique les méthodes aux estimations comme s'il s'agissait de paramètres directement observés ce qui nous conduit à travailler dans l'*espace de paramètres*. C'est l'approche qu'utilisent [81] pour décider si une molécule est discriminante ou pas, [82] pour classifier des signaux, ou [83] pour décider si une acquisition de cohorte fait ou ne fait pas partie d'une sous-cohorte donnée.

Approche Jointe Enfin, disposant des modélisations physique et probabiliste de la chaîne d'analyse, nous proposons d'utiliser la méthodologie des problèmes inverses. L'estimation de la classe peut être vue comme une estimation ponctuelle d'un paramètre discret, son estimation faisant appel à l'observation et aux modèles. On travaille alors dans un *espace joint*. C'est le cadre qui est présenté dans la suite de ce chapitre.

Fig. 5.1 Schémas pour trois types différents de classification : (a) l'approche *machine learning* (comparaison avec une donnée représentative), (b) l'approche séquentielle (avec des estimations intermédiaires) et (c) l'approche inversion présentée dans cette thèse (inversion du modèle physique, ici avec l'utilisation des méthodes statistiques bayésiennes).



Finalement, dans le cas où on dispose d'un modèle paramétrique et d'une modélisation probabiliste, s'impose un *troisième choix* car la structure du classifieur dépend également du choix des distributions pour chaque paramètre. Nous avons motivé le choix des distributions précédemment pour les paramètres autres que la classe et les paramètres de classe ; les distributions de ces derniers seront présentées dans la suite de ce chapitre.

5.2 État de l'art

La classification est un problème majeur dans le traitement de données et ses applications. Parmi les méthodes de classification, toutes ont le même but : estimer avec le moins d'erreur possible la classe de provenance d'une observation donnée. Cependant, la manière de s'y prendre diffère d'une méthode à l'autre, construisant des classifieurs différents. Dans ce qui suit, nous en exposons quatre qui sont issues de l'état de l'art « standard ». Elles nous serviront ensuite pour évaluer et comparer les performances du classifieur que nous introduirons après l'exposé de l'état de l'art.

5.2.1 Naïve Bayes

Le classifieur *Naïve Bayes* (NB) [84, Ch. 6.6.3] est une technique facile à comprendre et simple à mettre en œuvre, tout en étant relativement performante. Étant donnée une classe $C = c^m$, on suppose que les entrées (dans l'approche séquentielle, il s'agit des estimations de concentrations de chaque protéine) du vecteur $\hat{\boldsymbol{x}}$ sont indépendantes : $p(\hat{\boldsymbol{x}} \mid C = c^m) = \prod_{p=1}^{P} p(\hat{x}_p \mid C = c^m)$.

Le classifieur correspond finalement à l'estimateur Maximum A Posteriori

$$\psi_{NB}(\hat{\boldsymbol{x}}) = \underset{m=1,\ldots,M}{\operatorname{argmax}} \Pr_{NB}(C = c^m \mid \hat{\boldsymbol{x}}) \propto \Pr(C = c^m) \cdot \prod_{p=1}^{P} p(\hat{x}_p \mid c^m). \tag{5.1}$$

Ce classifieur trouve son avantage dans l'indépendance des entrées. Même si cette supposition est en général fausse, elle simplifie les développements : on peut se contenter de considérer plusieurs distributions mono-variées. Considérons par exemple le cas

gaussien, paramétré par la moyenne et la matrice de covariance. Au lieu de devoir apprendre une matrice de covariance pleine de taille $P \times P$ de $P(P-1)/2$ entrées distinctes, Naïve Bayes suppose une matrice de covariance diagonale et n'a ainsi besoin d'apprendre que des variances de chaque entrée, soit P paramètres.

5.2.2 Régression logistique

La régression logistique (LR) [84, Ch. 4.4] est issue du désir d'exprimer la probabilité d'un modèle ou d'une classe $\mathrm{c}^{\mathsf{m}} \in \mathcal{C}$ à l'aide d'une fonction linéaire en $\hat{\boldsymbol{x}}$ qui contient ici les estimations de concentration de protéines dans le contexte de l'approche séquentielle. Pour ce faire, on considère le logarithme des rapports des probabilités *a posteriori* des classes qui sera modélisé comme étant linéaire :

$$\log \frac{\Pr(C = \mathrm{c}^{\mathsf{m}} \mid \hat{\boldsymbol{x}})}{\Pr(C = \mathrm{c}^{\mathsf{M}} \mid \hat{\boldsymbol{x}})} = \alpha_{\mathsf{m}} + \beta_{\mathsf{m}}^{\mathrm{T}} \hat{\boldsymbol{x}}, \quad \forall \mathsf{m} = 1, \ldots, \mathsf{M}-1. \tag{5.2}$$

La probabilité de la classe M sert de facteur de normalisation ; son choix a été arbitraire, toute autre classe aurait pu faire l'affaire sans perte de généralité.

On isole la probabilité *a posteriori* d'une classe par prise de l'exponentiel :

$$\Pr(C = \mathrm{c}^{\mathsf{m}} \mid \hat{\boldsymbol{x}}) = \Pr(C = \mathrm{c}^{\mathsf{M}} \mid \hat{\boldsymbol{x}}) \exp(\alpha_{\mathsf{m}} + \beta_{\mathsf{m}}^{\mathrm{T}} \hat{\boldsymbol{x}}), \quad \mathsf{m} = 1, \ldots, \mathsf{M}-1. \tag{5.3}$$

Pour caractériser la probabilité de normalisation, on utilise le fait que la somme des probabilités est égale à l'unité :

$$\begin{aligned}
1 &= \sum_{\mathsf{m}=1}^{\mathsf{M}} \Pr(C = \mathrm{c}^{\mathsf{m}} \mid \hat{\boldsymbol{x}}) \\
1 &= \Pr(C = \mathrm{c}^{\mathsf{M}} \mid \hat{\boldsymbol{x}}) + \sum_{\mathsf{m}=1}^{\mathsf{M}-1} \Pr(C = \mathrm{c}^{\mathsf{m}} \mid \hat{\boldsymbol{x}}) \\
1 &= \Pr(C = \mathrm{c}^{\mathsf{M}} \mid \hat{\boldsymbol{x}}) + \Pr(C = \mathrm{c}^{\mathsf{M}} \mid \hat{\boldsymbol{x}}) \sum_{\mathsf{m}=1}^{\mathsf{M}-1} \exp(\alpha_{\mathsf{m}} + \beta_{\mathsf{m}}^{\mathrm{T}} \hat{\boldsymbol{x}}) \\
1 &= \Pr(C = \mathrm{c}^{\mathsf{M}} \mid \hat{\boldsymbol{x}}) \left(1 + \sum_{\mathsf{m}=1}^{\mathsf{M}-1} \exp(\alpha_{\mathsf{m}} + \beta_{\mathsf{m}}^{\mathrm{T}} \hat{\boldsymbol{x}}) \right)
\end{aligned}$$

ce qui fait finalement

$$\Pr(C = \mathrm{c}^{\mathsf{M}} \mid \hat{\boldsymbol{x}}) = \frac{1}{1 + \sum_{\mathsf{m}=1}^{\mathsf{M}-1} \exp(\alpha_{\mathsf{m}} + \beta_{\mathsf{m}}^{\mathrm{T}} \hat{\boldsymbol{x}})} \tag{5.4}$$

Compte tenu de la construction, les probabilités sont données par

$$\Pr_{\mathrm{LR}}(C = \mathrm{c}^{\mathsf{m}} \mid \hat{\boldsymbol{x}}) = \frac{\exp(\alpha_{\mathsf{m}} + \beta_{\mathsf{m}}^{\mathrm{T}} \hat{\boldsymbol{x}})}{1 + \sum_{r=1}^{\mathsf{M}-1} \exp(\alpha_r + \beta_r^{\mathrm{T}} \hat{\boldsymbol{x}})}, \quad \mathsf{m} = 1, \ldots, \mathsf{M}-1, \tag{5.5a}$$

$$\Pr_{\mathrm{LR}}(C = \mathrm{c}^{\mathsf{M}} \mid \hat{\boldsymbol{x}}) = \frac{1}{1 + \sum_{r=1}^{\mathsf{M}-1} \exp(\alpha_r + \beta_r^{\mathrm{T}} \hat{\boldsymbol{x}})}. \tag{5.5b}$$

La classification se fait à l'aide de l'estimateur du maximum de la loi, *i.e.*

$$\psi_{\mathrm{LR}}(\hat{\boldsymbol{x}}) = \operatorname*{argmax}_{\mathsf{m}=1,\ldots,\mathsf{M}} \Pr_{\mathrm{LR}}(C = \mathrm{c}^{\mathsf{m}} \mid \hat{\boldsymbol{x}}). \tag{5.6}$$

Notons cependant que les estimateurs Naïve Bayes et Régression logistique ne maximisent pas les mêmes lois : alors que le NB impose l'indépendance des entrées de la quantité à classifier, la LR impose plutôt la linéarité de l'expression de la probabilité de la classe.

La remarque suivante propose une manière d'estimer les paramètres α_{m} et β_{m} ; la référence citée ci-dessus en expose d'autres en s'inspirant de méthodes différentes.

Remarque ⟨5.1⟩ (Régression logistique et estimation bayésienne) *Dans cette remarque, nous allons montrer une équivalence entre l'estimation bayésienne de la classe et la régression logistique* sous certaines conditions restrictives. *Elles seront introduites au fur et à mesure. Cette équivalence nous permettra de proposer une estimation des paramètres de la régression logistique. Nous étudierons le cas binaire uniquement ; le cas N-aire s'obtient facilement d'une manière analogue.*

Soit $\mathcal{C} = \{\mathrm{c}^0, \mathrm{c}^1\}$ *l'ensemble des classes auxquelles est associée la distribution* a priori $\Pr(C = \mathrm{c}^0) = p_0 = 1 - p_1 = 1 - \Pr(C = \mathrm{c}^1)$.

Ensuite, nous introduisons la première restriction *: soit* $\mathrm{p}(\boldsymbol{x} \mid \mathrm{c}^{\mathsf{m}}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_{\mathsf{m}}, \boldsymbol{\Gamma}_{\mathsf{m}})$ *une loi normale pour* $\mathsf{m} \in \{0, 1\}$.

La loi a posteriori *pour la classe* c^0 *est donnée par*

$$\begin{aligned}\Pr(\mathrm{c}^0 \mid \boldsymbol{x}) &= \frac{p_0 \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_0, \boldsymbol{\Gamma}_0)}{p_0 \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_0, \boldsymbol{\Gamma}_0) + p_1 \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_1, \boldsymbol{\Gamma}_0)} = \frac{\frac{\mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_0, \boldsymbol{\Gamma}_0)}{\mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_1, \boldsymbol{\Gamma}_1)}}{\underbrace{\frac{\mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_0, \boldsymbol{\Gamma}_0)}{\mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_1, \boldsymbol{\Gamma}_1)}}_{=:\mathsf{E}(\boldsymbol{m}_0, \boldsymbol{m}_1, \boldsymbol{\Gamma}_0, \boldsymbol{\Gamma}_1)} + \frac{p_1}{p_0}} \\ &= \frac{\mathsf{E}(\boldsymbol{m}_0, \boldsymbol{m}_1, \boldsymbol{\Gamma}_0, \boldsymbol{\Gamma}_1)}{\mathsf{E}(\boldsymbol{m}_0, \boldsymbol{m}_1, \boldsymbol{\Gamma}_0, \boldsymbol{\Gamma}_1) + \frac{p_1}{p_0}}.\end{aligned}$$

La fonction $\mathsf{E}(\boldsymbol{m}_0, \boldsymbol{m}_1, \boldsymbol{\Gamma}_0, \boldsymbol{\Gamma}_1)$ *est un rapport de deux lois normales qui se développe de la façon suivante :*

$$\mathsf{E}(\boldsymbol{m}_0, \boldsymbol{m}_1, \boldsymbol{\Gamma}_0, \boldsymbol{\Gamma}_1) = \left(\frac{|\boldsymbol{\Gamma}_0|}{|\boldsymbol{\Gamma}_1|}\right)^{1/2} \exp\left(-\tfrac{1}{2}\left[\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_1)\boldsymbol{x} - 2(\boldsymbol{m}_0^{\mathrm{T}}\boldsymbol{\Gamma}_0 - \boldsymbol{m}_1^{\mathrm{T}}\boldsymbol{\Gamma}_1)\boldsymbol{x} + \mathrm{const}\right]\right).$$

L'argument de l'exponentielle définit la conique de séparation entre les deux classes. Pour $\boldsymbol{\Gamma}_0 \neq \boldsymbol{\Gamma}_1$*, l'hyperplan est parabolique puisque l'argument est un polynôme de degré* 2.

La deuxième restriction *s'effectue au niveau des précisions des lois normales. On suppose des lois normales homoscédastiques. Soit alors maintenant* $\boldsymbol{\Gamma}_0 = \boldsymbol{\Gamma}_1$*. L'expression de* E *se simplifie et on obtient*

$$\mathsf{E}(\boldsymbol{m}_0, \boldsymbol{m}_1, \boldsymbol{\Gamma}) = \exp\big(\underbrace{(\boldsymbol{m}_0 - \boldsymbol{m}_1)^{\mathrm{T}}\boldsymbol{\Gamma}}_{=\beta^{\mathrm{T}}}\boldsymbol{x} + \underbrace{\mathrm{const}}_{=\alpha}\big) = \exp(\beta^{\mathrm{T}}\boldsymbol{x} + \alpha).$$

La conique de séparation des classes n'est plus qu'une droite, définie par les paramètres (α, β) *[85]. La probabilité* a posteriori *pour* c^0 *s'écrit alors*

$$\Pr(\mathrm{c}^0 \mid \boldsymbol{x}) = \frac{\exp(\beta^{\mathrm{T}}\boldsymbol{x} + \alpha)}{\exp(\beta^{\mathrm{T}}\boldsymbol{x} + \alpha) + p_1/p_0}.$$

Nous avons ainsi déduit d'une classification bayésienne sur des données directes l'expression de la régression logistique à distribution a priori *non uniforme sur les classes. Si* $p_0 = p_1$ *ce qui correspond à la* troisième restriction*, alors on retrouve l'Éqn. (5.5). Il en résulte de plus que l'hyperplan séparant les classes dans la régression logistique est une droite dont les coefficients* (α, β) *peuvent s'obtenir suivant la construction décrite ci-avant.*

5.2.3 k-means

L'approche *k-means* [84, Sect. 14.3.6] est une méthode de partitionnement, *i.e.* d'étiquetage non supervisé, qui tente de partitionner N observations dans M partitions[1], le nombre de partitions étant choisi *a priori*. Les associations de partition aux observations se font globalement en minimisant le critère de somme des carrés à l'intérieur de chaque partition (*Within Partition Squared Sums*)

$$\mathbb{J} = \sum_{\mathsf{m}=1}^{\mathsf{M}} \sum_{x_j \in \mathsf{c}^{\mathsf{m}}} \left\| x_j - \boldsymbol{m}_{\mathsf{m}} \right\|^2 ;$$

ainsi, l'observation x_j appartient à la partition c^{m} dont le centre $\boldsymbol{m}_{\mathsf{m}}$ est le plus proche. (L'utilisation d'une autre distance que la distance euclidienne est tout à fait autorisée et peut trouver son intérêt dans la résolution de problèmes vivant dans un espace non euclidien.)

Un certain nombre d'extensions ont été proposées, certaines pour diminuer l'impact de l'initialisation ou pour atteindre une meilleure convergence. Pour pallier l'inflexibilité du nombre de partitions, la méthode « *Sum Of Norms* » [86], s'inspirant des k-means, en propose une détermination automatique, contrairement aux k-means. Pour cela, la méthode travaille avec autant de partitions qu'il y a d'observations. Ensuite, grâce à un terme qui pénalise la norme de la différence des centres des partitions qui s'ajoute au terme à minimiser, certains centres se confondent. Finalement, toutes observations dont les centres sont égaux appartiennent à une même partition.

Il existe également une variante des k-means s'inscrivant dans les méthodes de la logique floue qui n'assigne pas une et une seule partition à une observation, mais plusieurs avec un certain degré de confiance. Nous introduisons cette variante dans la section suivante.

5.2.4 Fuzzy c-means

La méthode des *fuzzy c-means* [87] est une extension de la méthode *k-means* à la logique floue. Toutes les deux sont des méthodes de partitionnement non-supervisé, visant à estimer le nombre de classes, leurs centroïdes, et l'appartenance aux classes détectées. Contrairement aux *k-means* où l'appartenance n'est possible qu'à une seule classe, les *fuzzy c-means* (FCM) permettent l'appartenance à plusieurs classes. Pour cela, on minimise un critère quadratique pondéré,

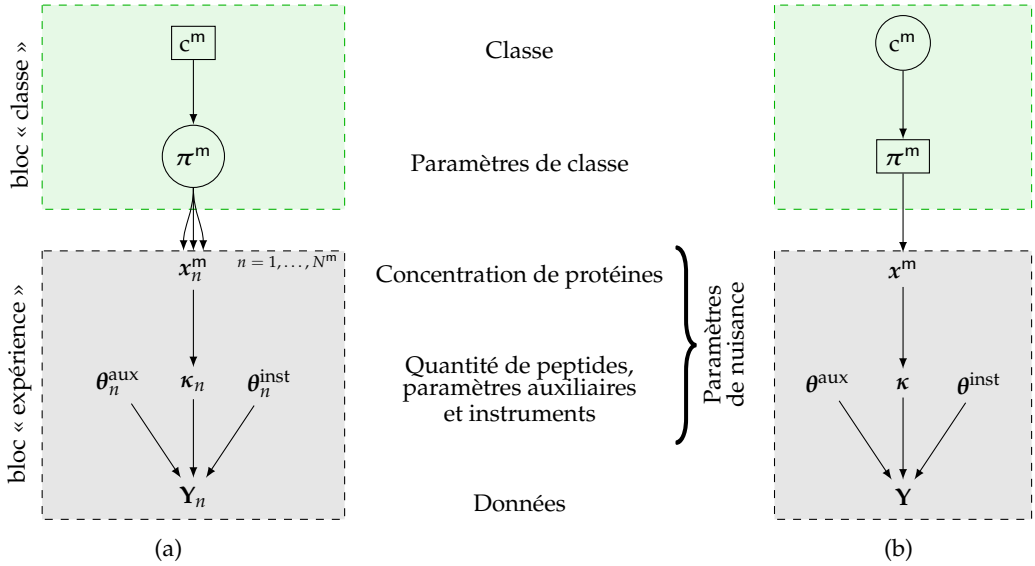
$$\sum_{n=1}^{N} \sum_{\mathsf{m}=1}^{\mathsf{M}} (\mu_n^{\mathsf{m}})^f \, \|\boldsymbol{x}_n - m_{\mathsf{m}}\|_{\mathbf{A}}^2$$

où f est le degré de flou (*level of fuzziness*), m_{m} le centroïde de la classe m, et $\|\cdot\|_{\mathbf{A}}$ une norme matricielle pondérée par la matrice $\mathbf{A}$[2]. Ainsi, on calcule pour un échantillon $n = 1, \ldots, N$ (une estimation de concentration de protéines) un *degré d'appartenance* μ_n^{m} à la classe $\mathsf{c}^{\mathsf{m}} \in \mathcal{C}$ dont la valeur est comprise entre 0 et 1, la somme des degrés pour un échantillon donné égalant 1. Il est important de remarquer que cette valeur **n'est pas** une probabilité et ne doit pas être interprétée comme telle.

L'extension au cas de la classification avec un apprentissage supervisé est immédiate. Soient les échantillons indicés par $n = 1, \ldots, N$ les échantillons d'apprentissage. Il suffit d'imposer $\mu^{\mathsf{m},n} = \delta(\mathsf{c}_n^\star, \mathsf{c}^{\mathsf{m}})$ pour l'échantillon n qui vaut 1 si la classe c^{m} correspond à

1. Classiquement, la méthode s'appelle *k-means* car le nombre de partitions est k. Dans la suite, nous utilisons toujours M partitions, sans changer le nom « standardisé » de cette méthode en M-*means*.
2. Un choix récurrent pour ces paramètres est $f = 2$ et $\mathbf{A} = \mathbf{I}$ la matrice identité.

Fig. 5.2 Diagramme hiérarchique de la classification. (a) pour l'apprentissage, (b) pour le classement. Les variables emboîtées sont connues, les variables encerclées sont d'intérêt.



la classe vraie $c_n^\star$ de l'échantillon, et 0 sinon. L'ensemble d'échantillons d'apprentissage à degré d'appartenance fixe et d'échantillons à classer est ensuite partitionné comme décrit ci-dessus.

5.2.5 Bilan

Les méthodes de classification exposées ci-dessus présentent certains défauts. Dans les approches *Machine Learning*, la recherche des caractéristiques différentiantes peut être difficile compte tenu de la nature des signaux chromato-spectrométriques et les variabilités impliqués. De plus, le modèle direct qui nous avons mis en place n'est pas considéré alors qu'il nous apporte une information qu'il est crucial d'utiliser.

L'approche séquentielle est utilisée couramment, notamment Naïve Bayes et – en biostatistiques – la régression logistique. Cependant, on travaille sur des estimations pour faire une deuxième estimation. La première comporte déjà un choix sur l'estimateur ponctuel (bayésien ou non) et néglige l'information sur la variabilité et sur l'incertitude.

Cette thèse propose d'aller plus loin en fusionnant les deux en travaillant dans un *espace joint* sans estimation séquentielle (Fig. 5.1(c)). L'estimation est faite en inversant le modèle direct de la chaîne d'analyse qui explique la génération des données par les « causes » (les paramètres d'intérêt) en passant par les « instruments » (paramètres de nuisance) jusqu'aux « conséquences » (observations).

5.3 Apprendre

L'objet de la présente section est l'estimation des paramètres des classes π. Pour cela, nous identifions les paramètres d'une classe avec les paramètres de la distribution de concentration des protéines. Sous les conditions d'indépendance que nous imposons, nous montrons que l'apprentissage est séparable par rapport aux classes ce qui nous permet de faire des apprentissages mono-classe. Nous exprimons l'estimateur qui sera calculé par l'algorithme MCMC dont nous donnerons le schéma. À la fin de cette section, nous donnons des résultats et en évaluons les performances sur des données simulées.

La situation est la suivante :
- Les données d'apprentissage, regroupées dans l'ensemble $\mathcal{T} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\} \equiv \{\mathbf{Y}_{1:N}\}$, proviennent d'une cohorte de taille N ;
- à l'intérieur de la cohorte, il y a M sous-cohortes ;
- chaque sous-cohorte $\mathcal{T}^{\mathsf{m}} = \{\mathbf{Y}_n \mid \mathbf{Y}_n \in \mathrm{c}^{\mathsf{m}},\ n = 1, \ldots, N\}$ est de taille N^{m} pour $\mathsf{m} = 1, \ldots, \mathsf{M}$ et $\sum_{\mathsf{m}=1}^{\mathsf{M}} N^{\mathsf{m}} = N$;[3]
- l'ensemble des sous-cohortes $\mathcal{T}^{1:\mathsf{M}}$ est une partition de $\mathcal{T}$;
- les membres de la sous-cohorte $\mathcal{T}^{\mathsf{m}}$ sont notés $\mathbf{Y}_n^{\mathsf{m}}$, et la sous-cohorte elle-même peut s'écrire $\mathcal{T}^{\mathsf{m}} = \{\mathbf{Y}_{1:N^{\mathsf{m}}}^{\mathsf{m}}\}$
- à chaque classe c^{m} est associé un paramètre de classe $\boldsymbol{\pi}^{\mathsf{m}}$ qui correspond aux hyperparamètres de la distribution de $\boldsymbol{x} \mid \mathrm{c}^{\mathsf{m}}$.

La situation est représentée schématiquement dans la Fig. 5.2(a).

Remarque ⟨5.2⟩ *Dans la récolte de données cliniques, il se peut qu'une attribution de classe à un individu d'apprentissage soit erronée. C'est le cas par exemple quand la cohorte d'individus contrôle est issue d'un centre de don de sang. On peut tomber sur un individu qui – à son insu – est dans un stade préliminaire d'une maladie considérée. On parle alors d'« étiquette aberrante ».*

Nous supposons que ce n'est pas le cas ici. *Cependant, si on voulait en tenir compte, la modélisation serait à revoir en y ajoutant notamment un indicateur « erreur d'étiquetage », soit pour exclure l'individu, soit pour lui attribuer une autre classe. Une des méthodes correctrices couramment utilisées, on compte évidemment la validation croisée où on cherche à reclasser l'individu, l'apprentissage étant fait sur la cohorte sans l'individu en question. Pour un nombre N grand, l'effort est très important au vu des N apprentissages. Au lieu de considérer un individu, on peut en considérer V pour en faire des paquets de reclassement avec l'apprentissage sur l'ensemble de la cohorte sans les V individus choisis. En dehors de cela, [88] étudie cette question en associant un poids correcteur à chaque concentration. Ainsi, la concentration est ramenée vers la bonne classe.*

Ceci étant dit, le cadre bayésien fournit tous les outils pour décider si une étiquette est aberrante ou non. En ajoutant un paramètre d'indicateur, on peut calculer la probabilité de remise en cause de l'étiquette (compte tenu de la cohorte d'apprentissage) ; le cas échéant une ré-attribution d'étiquette est possible, la méthode s'inscrivant ainsi dans l'apprentissage semi-supervisé (ou apprentissage à supervision partielle).

5.3.1 Séparation naturelle des classes

Chaque donnée d'apprentissage est acquise indépendamment des autres ; dans nos termes, les vraisemblances des acquisitions se séparent. De plus, les paramètres étant également indépendants d'une expérience à l'autre, on modélise l'indépendance *a priori*. Ainsi, la distribution jointe des données de la cohorte $\mathcal{T}$ se décompose en un produit de plusieurs distributions. Regroupons momentanément les sous-cohortes. Nous obtenons donc M distributions jointes au niveau des sous-cohortes $\mathcal{T}^{\mathsf{m}}$:

$$\mathrm{p}\left(\mathcal{T}, \boldsymbol{\pi}^{1:\mathsf{M}}, \left[\boldsymbol{x}_{1:N^{\mathsf{M}}}^{\mathsf{m}}\right]^{1:\mathsf{M}}, \left[\boldsymbol{\kappa}_{1:N^{\mathsf{M}}}\right]^{1:\mathsf{M}}, \left[\boldsymbol{\theta}_{1:N^{\mathsf{M}}}^{\mathrm{inst}}\right]^{1:\mathsf{M}}\right) = \prod_{\mathsf{m}=1}^{\mathsf{M}} \mathrm{p}\left(\mathcal{T}^{\mathsf{m}}, \boldsymbol{\pi}^{\mathsf{m}}, (\boldsymbol{x}^{\mathsf{m}})_{1:N^{\mathsf{m}}}, (\boldsymbol{\kappa})_{1:N^{\mathsf{m}}}, (\boldsymbol{\theta}^{\mathrm{inst}})_{1:N^{\mathsf{m}}}\right). \quad (5.7)$$

À l'intérieur de chaque regroupement, la distribution jointe des données de la sous-cohorte $\mathcal{T}^{\mathsf{m}}$ et des paramètres se décompose en un produit de $N^{\mathsf{m}} + 1$ distributions :

3. L'opérateur $\in$ se lit ici « appartient à la classe », « dont l'étiquette est ».

$$\mathrm{p}\left(\mathcal{T}^{\mathsf{m}}, \boldsymbol{\pi}^{\mathsf{m}}, (\boldsymbol{x}^{\mathsf{m}})_{1:N^{\mathsf{m}}}, (\boldsymbol{\kappa})_{1:N^{\mathsf{m}}}, (\boldsymbol{\theta}^{\mathrm{inst}})_{1:N^{\mathsf{m}}}\right) = \mathrm{p}(\boldsymbol{\pi}^{\mathsf{m}}) \cdot \left[\prod_{n=1}^{N^{\mathsf{m}}} \mathrm{p}\left(\mathbf{Y}_n^{\mathsf{m}}, \boldsymbol{x}_n^{\mathsf{m}}, \boldsymbol{\kappa}_n, (\boldsymbol{\theta}^{\mathrm{inst}})_n\right)\right]. \tag{5.8}$$

On peut déduire des deux équations précédentes que l'apprentissage global sur l'ensemble de la cohorte de taille N se sépare automatiquement en M sous-apprentissages mono-classe indépendantes ce qui peut sembler intuitif puisqu'il s'agit d'un apprentissage supervisé. C'est la raison pour laquelle nous allons mettre le focus sur l'apprentissage à l'intérieur d'une seule classe sans perte de généralité ; nous omettrons donc dans la suite de la section l'indice de la classe.

5.3.2 Modèle direct et loi jointe

Grâce à l'indépendance conditionnelle des expériences et à la structure hiérarchique du problème direct, la loi jointe des paramètres d'une classe est composée d'un facteur « classe » et de N facteurs « expériences » :

$$\begin{aligned}\mathrm{p}\left(\mathcal{T}, \boldsymbol{\pi}, \boldsymbol{x}_{1:N}, \boldsymbol{\kappa}_{1:N}, \boldsymbol{\theta}_{1:N}^{\mathrm{inst}}\right) &= \mathrm{p}(\boldsymbol{\pi}) \cdot \left[\prod_{n=1}^{N} \mathrm{p}\left(\mathbf{Y}_n, \boldsymbol{x}_n, \boldsymbol{\kappa}_n, \boldsymbol{\theta}_n^{\mathrm{inst}}\right)\right] \\ &= \underbrace{\mathrm{p}(\boldsymbol{\pi})}_{\text{classe}} \cdot \prod_{n=1}^{N} \underbrace{\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\pi})\, \mathrm{p}(\boldsymbol{\kappa} \mid \boldsymbol{x})\, \mathrm{p}(\boldsymbol{\theta}^{\mathrm{inst}})\, \mathrm{p}(\mathbf{Y} \mid \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}})}_{\text{expérience}}. \end{aligned} \tag{5.9}$$

On voit bien qu'au conditionnement par $\boldsymbol{\pi}$ près, chaque facteur « expérience » pour un n donné correspond à la loi jointe du problème de quantification comme nous l'avons introduit dans le Ch. 4.

Il convient maintenant de choisir une loi *a priori* pour les paramètres de la classe $\boldsymbol{\pi}$. La distribution de la concentration des protéines étant normale, les paramètres de la classe sont naturellement la moyenne et la précision de cette loi, *i.e.* les hyperparamètres de la distribution de la concentration : $\boldsymbol{\pi} = [\boldsymbol{m}, \boldsymbol{\Gamma}]$. Suite à la nature des paramètres et au vu d'une conjugaison de loi *a priori* par la vraisemblance hiérarchique associée, nous choisissons la loi **normale-wishartienne**[xix] $\mathcal{NW}(\boldsymbol{\pi}; \boldsymbol{\mu}, \boldsymbol{\Lambda}, \eta, \nu)$ dont les paramètres s'interprètent comme suit. Les paramètres $\boldsymbol{\Lambda}$ et ν décrivent la matrice d'échelle et les degrés de liberté pour la distribution *a priori* wishartienne sur le paramètre de précision $\boldsymbol{\Gamma}$. La moyenne *a priori* de l'échantillon utilisé est $\boldsymbol{\mu}$ et la taille *a priori*– qui peut être utilisée pour nuancer le poids que doit prendre l'apport de l'*a priori*– est donnée par η. Le choix d'injecter le moins d'information possible se fait en proposons une matrice d'échelle de déterminant et un paramètre de taille *a priori* proches de 0.

Remarque ⟨5.3⟩ *Cette écriture nous permet de « mettre à jour » l'apprentissage avec des données supplémentaires acquises ultérieurement à une campagne expérimentale sans devoir faire un nouvel apprentissage complet.*

Partons des paramètres de la classe que nous avons déjà estimés à partir d'une cohorte d'effectif η individus. À l'aide de N nouvelles acquisitions, la mise à jour s'effectue en prenant comme paramètres a priori *les suivants : la moyenne et la précision estimées lors du premier apprentissage pour $\boldsymbol{\mu}$ et $\boldsymbol{\Lambda}$, le nombre d'individus utilisé pour cette estimation η, et le nombre de degrés de liberté ν.*

La loi normale-wishartienne qui est le produit entre une loi normale et une loi wi-

shartienne a une forme explicite et s'écrit

$$\mathcal{NW}(\boldsymbol{m}, \boldsymbol{\Gamma}; \boldsymbol{\mu}, \boldsymbol{\Lambda}, \eta, \nu) = \frac{|\boldsymbol{\Lambda}|^{-\nu/2}}{2^{\nu P/2}(2\pi)^P \Gamma_P(\nu/2)} |\boldsymbol{\Gamma}|^{(\nu-P)/2} \cdot \exp\left(-\tfrac{1}{2}\left[\operatorname{tr}\left(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\right) + \eta\,(\boldsymbol{\mu}-\boldsymbol{m})^{\mathrm{T}}\,\boldsymbol{\Gamma}\,(\boldsymbol{\mu}-\boldsymbol{m})\right]\right) \quad (5.10)$$

où P est la dimension correspondant au nombre de protéines, $\Gamma_P(\cdot)$ est la fonction Gamma P-dimensionnelle, $\Gamma_P(x) = \pi^{P(P-1)/4}\prod_{p=1}^{P}\Gamma\left(x + (1-p)/2\right)$, et $\operatorname{tr}(\cdot)$ l'opérateur trace d'une matrice carrée.

Remarque ⟨5.4⟩ (Wishart et Riemann) *La distribution wishartienne est souvent choisie pour l'échantillonnage d'une matrice de précision, la distribution inverse wishartienne pour les matrices de covariance. Pourquoi ce choix est-il raisonnable ? À l'intérieur de l'argument de la fonction exponentielle de l'Éqn. (5.10), la trace d'un produit matriciel entre A et l'inverse de B est calculée. La référence [89] constate que les matrices de précision et de covariances vivent dans une géométrie riemanienne dont la fonction de distance entre deux matrices A et B est calculée à partir la somme du logarithme des valeurs propres du produit de A et l'inverse de B :*

$$\operatorname{dist}_R(A, B) = \left(\sum_n (\ln \lambda_n(A, B))^2\right)^{1/2}.$$

La trace est invariante par changement de base, d'où $\operatorname{tr}(AB^{-1}) = \operatorname{tr}(S^{-1}DS) = \operatorname{tr}(SS^{-1}D) = \operatorname{tr}(D) = \sum_n \lambda_n(A, B)$, où D est la matrice diagonale du produit AB^{-1}, composée des valeurs propres de ce dernier.

Le logarithme étant monotone et croissant, nous pouvons donc conclure que maximiser la distribution wishartienne d'une matrice par rapport à la matrice d'échelle correspond à minimiser la distance riemannienne entre ces deux matrices.

5.3.3 Expression de l'estimateur

L'objectif est d'estimer les paramètres de la classe considérée. Soit $\psi^{\mathrm{a}} : \Omega_{\mathbf{Y}}^N \to \Omega_{\boldsymbol{\pi}}$ un estimateur qui associe à une cohorte d'apprentissage $\mathcal{T}$ de taille N l'estimation $\psi^{\mathrm{a}}(\mathcal{T}) = \hat{\boldsymbol{\pi}}$ de $\boldsymbol{\pi}$. Pour exprimer l'erreur quadratique de l'estimation $\psi(\mathcal{T})$ par rapport à une valeur vraie notée $\boldsymbol{\pi}^\star$, nous transformons le couple $(\boldsymbol{m}, \boldsymbol{\Gamma}) \in \mathbb{R}_+^P \times \mathbb{R}^{P\times P}$ en un vecteur de taille $\mathbb{R}^{P+P^2}$ en concaténant le vecteur de la moyenne et la vectorisation de la précision. De par ce qui suit, nous pouvons le noter également par $\boldsymbol{\pi}$. Il est immédiat de voir que la norme-2 carrée de ce vecteur est la somme de la norme-2 carrée de la moyenne et la norme carrée de la précision au sens de Frobenius :

$$\|\boldsymbol{\pi}\|^2 = \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{\pi} = \sum_{p=1}^{P+P^2} \pi_p^2 = \sum_{p=1}^{P} \boldsymbol{m}_p^2 + \sum_{u=1}^{P}\sum_{v=1}^{P} [\boldsymbol{\Gamma}]_{uv}^2 = \|\boldsymbol{m}\|_2^2 + \|\boldsymbol{\Gamma}\|_{\mathrm{F}}^2.$$

L'estimateur bayésien attaché à la fonction de coût correspondant à l'erreur quadratique est, comme nous l'avons dit au Ch. 2, l'estimateur Espérance A Posteriori. Il minimise le risque bayésien, *i.e.* le coût moyen pris sous la loi jointe des paramètres et des données. Son expression dans ce cas s'écrit

$$\hat{\boldsymbol{\pi}}_{\mathrm{EAP}} = \psi_{a,\,\mathrm{EAP}}(\mathcal{T}) = \mathbb{E}_{\boldsymbol{\Pi}|\mathcal{T}}(\boldsymbol{\pi} \mid \mathcal{T}) = \int_{\Omega_{\boldsymbol{\pi}}} \boldsymbol{\pi}\, \mathrm{p}(\boldsymbol{\pi} \mid \mathcal{T})\, \mathrm{d}\boldsymbol{\pi}. \quad (5.11)$$

L'estimateur appelle implicitement une marginalisation des paramètres de nuisance $\left[\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}}\right]_n$ de chaque expérience $n = 1, \ldots, N$:

$$\psi_{\mathrm{EAP}}(\mathcal{T}) = \int_{\Omega_{\boldsymbol{\pi}}} \int_{\Omega_{\boldsymbol{x},\boldsymbol{\kappa},\boldsymbol{\theta}^{\mathrm{inst}}}} \cdots \int \boldsymbol{\pi}\, \mathrm{p}(\boldsymbol{\pi}, \boldsymbol{x}_{1:N}, \boldsymbol{\kappa}_{1:N}, \boldsymbol{\theta}_{1:N}^{\mathrm{inst}} \mid \mathcal{T})\, \mathrm{d}(\boldsymbol{x}_{1:N}, \boldsymbol{\kappa}_{1:N}, \boldsymbol{\theta}_{1:N}^{\mathrm{inst}})\, \mathrm{d}\boldsymbol{\pi}. \quad (5.12)$$

Le calcul analytique de la marginalisation, puis de l'intégration du paramètre $\boldsymbol{\pi}$ par rapport à la mesure de probabilité $\mathrm{p}(\boldsymbol{\pi} \mid \mathcal{T})$ est impossible : le nombre de paramètres impliqués est élevé, la dimension des données est importante, et la loi *a posteriori* n'a pas de forme « standard ». Cependant, en utilisant un échantillonnage stochastique, nous pouvons transformer ce problème d'intégration en un problème d'échantillonnage qui est plus facile à résoudre. Pour cela, nous avons recours aux méthodes MCMC (Sect. 2.9.1).

5.3.4 Mise en œuvre de l'inversion

Nous proposons de calculer l'estimateur EAP du couple $(\boldsymbol{m}, \boldsymbol{\Gamma}) = \boldsymbol{\pi}$ par échantillonnage stochastique. Pour cela, l'approche MCMC associée à la structure de Gibbs est un moyen très efficace. Le calcul de l'EAP revient à moyenner les échantillons *a posteriori* des paramètres à estimer :

$$\left[\hat{\boldsymbol{m}}, \hat{\boldsymbol{\Gamma}}\right] = \frac{1}{K} \sum_{k=K_0+1}^{K_0+K} \left[\boldsymbol{m}^{(k)}, \boldsymbol{\Gamma}^{(k)}\right] \equiv \frac{1}{K} \sum_{k=K_0+1}^{K_0+K} \left[(\boldsymbol{\pi})^{(k)}\right]. \tag{5.13}$$

Comme nous l'avons déjà fait pour la quantification, nous allons d'abord identifier la structure qui permet d'échantillonner efficacement avec un échantillonneur de Gibbs. Ensuite, nous déduirons les lois *a posteriori* conditionnelles. Enfin, si l'échantillonnage des paramètres n'est pas explicite, nous déciderons d'une méthode d'échantillonnage.

Lois a posteriori conditionnelles pour les paramètres de classe

Nous avons factorisé la loi jointe Éqn. (5.9) en *classe* et *expérience*. Pour chaque expérience, nous retrouvons les paramètres que nous avons déjà introduits dans le Ch. 4. Ceci permet alors de reconsidérer les lois que nous y avons justifiées et déduites : la loi *a posteriori* conditionnelle pour $\boldsymbol{\theta}_n^{\text{inst}}$, $\boldsymbol{\kappa}_n$ et la vraisemblance totale $\mathrm{p}(\mathbf{Y}_n \mid \boldsymbol{\kappa}_n, \boldsymbol{\theta}_n^{\text{inst}})$ par expérience sont les identiques à celles de la partie « Quantification ».

Nous devons cependant mettre à jour la loi pour la concentration des protéines. Sa loi *a priori* étant régie par les paramètres de la classe, $\mathrm{p}(\boldsymbol{x}_n \mid \boldsymbol{\pi}) = \mathrm{p}(\boldsymbol{x}_n \mid \boldsymbol{m}, \boldsymbol{\Gamma}) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{m}, \boldsymbol{\Gamma})$, nous retrouvons cette dépendance hiérarchique dans la loi *a posteriori* $\mathrm{p}(\boldsymbol{x}_n \mid \boldsymbol{\pi}, \boldsymbol{\kappa}_n) \propto \mathrm{p}(\boldsymbol{x}_n \mid \boldsymbol{\pi})\, \mathrm{p}(\boldsymbol{\kappa}_n \mid \boldsymbol{x}_n)$. Il s'agit de deux lois normales où $\boldsymbol{x}_n$ est en dépendance linéaire par rapport à $\boldsymbol{\kappa}_n$. Les lois sont conjuguées, et la loi *a posteriori* $\mathrm{p}(\boldsymbol{x}_n \mid \boldsymbol{\pi}, \boldsymbol{\kappa}_n)$ est donc également une loi normale $\mathcal{N}(\boldsymbol{x}_n; \boldsymbol{m}_{\boldsymbol{x}_n}^{\text{post}}, \boldsymbol{\Gamma}_{\boldsymbol{x}_n}^{\text{post}})$.

Quant aux paramètres de classe $\boldsymbol{\pi}$ et auxquels nous avons attribué conjointement une loi normale-wishartienne, la vraisemblance conditionnelle $\mathrm{p}(\boldsymbol{x}_{1:N} \mid \boldsymbol{\pi}) = \prod_{n=1}^{N} \mathrm{p}(\boldsymbol{x}_n \mid \boldsymbol{\pi})$ est normale et conjugue la loi normale-wishartienne ([3, Sect. 3.6], voir aussi Ann. B.3 pour la justification et pour le calcul des paramètres *a posteriori*). La loi *a posteriori* conditionnelle pour $\boldsymbol{\pi}$ a donc la forme

$$\mathrm{p}(\boldsymbol{\pi} \mid \boldsymbol{x}_{1:N}) = \mathcal{NW}(\boldsymbol{\pi}; \boldsymbol{\mu}^{\text{post}}, \boldsymbol{\Lambda}^{\text{post}}, \eta^{\text{post}}, \nu^{\text{post}}) \tag{5.14}$$

avec comme paramètres *a posteriori* $(\boldsymbol{\mu}^{\text{post}}, \boldsymbol{\Lambda}^{\text{post}}, \eta^{\text{post}}, \nu^{\text{post}})$.

Nous résumons les lois *a posteriori* des deux groupes de paramètres $\boldsymbol{x}_n$ et $\boldsymbol{\pi}$. Pour les autres paramètres, nous pouvons nous référer à la Sect. 4.3.1, page 72, modulo l'indice n de l'expérience.

(1) concentration des protéines cibles $\boldsymbol{x}_n$ de l'expérience n

<u>a priori :</u> $\mathrm{p}(\boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{m}, \boldsymbol{\Gamma})$
<u>vraisemblance associée :</u> $\mathrm{p}(\boldsymbol{\kappa}_n \mid \boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{\kappa}_n; \mathbf{D}\boldsymbol{x}_n, \boldsymbol{\Gamma}_{\boldsymbol{\kappa}})$
<u>conjugaison :</u> oui car la vraisemblance de $\boldsymbol{x}_n$ a la même forme que son *a priori*

<u>a posteriori :</u> $\mathrm{p}(\boldsymbol{x}_n \mid \boldsymbol{m}, \boldsymbol{\Gamma}, \boldsymbol{\kappa}_n) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_{\boldsymbol{x}_n}^{\text{post}}, \boldsymbol{\Gamma}_{\boldsymbol{x}_n}^{\text{post}})$
- $\boldsymbol{m}_{\boldsymbol{x}_n}^{\text{post}} = (\boldsymbol{\Gamma}_n^{\text{post}})^{-1}(\boldsymbol{\Gamma}\boldsymbol{m} + \mathbf{D}^{\mathrm{T}}\boldsymbol{\Gamma}_{\boldsymbol{\kappa}}\boldsymbol{\kappa}_n)$,
- $\boldsymbol{\Gamma}_{\boldsymbol{x}_n}^{\text{post}} = \boldsymbol{\Gamma} + \mathbf{D}^{\mathrm{T}}\boldsymbol{\Gamma}_{\boldsymbol{\kappa}}\mathbf{D}$;

<u>échantillonnage :</u> explicite

(2) paramètres de la classe $(\boldsymbol{m}, \boldsymbol{\Gamma})$

<u>a priori :</u> $\mathrm{p}(\boldsymbol{m}, \boldsymbol{\Gamma}) = \mathcal{NW}(\boldsymbol{m}, \boldsymbol{\Gamma}; \boldsymbol{\mu}, \boldsymbol{\Lambda}, \eta, \nu)$

<u>vraisemblance associée :</u> $\mathrm{p}(\boldsymbol{x}_n \mid \boldsymbol{m}, \boldsymbol{\Gamma}) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{m}, \boldsymbol{\Gamma})$

<u>conjugaison :</u> oui, voir Ann. B.3

<u>a posteriori :</u> $\mathrm{p}(\boldsymbol{m}, \boldsymbol{\Gamma} \mid \boldsymbol{x}_{1:N}) = \mathcal{NW}(\boldsymbol{m}, \boldsymbol{\Gamma}; \boldsymbol{\mu}^{\text{post}}, \boldsymbol{\Lambda}^{\text{post}}, \eta^{\text{post}}, \nu^{\text{post}})$
- $\boldsymbol{\mu}^{\text{post}} = \frac{\eta\boldsymbol{\mu} + N\bar{\boldsymbol{x}}}{\eta + N}$,
- $\boldsymbol{\Lambda}^{\text{post}} = \left[\boldsymbol{\Lambda}^{-1} + \frac{N\eta}{\eta^{\text{post}}}(\boldsymbol{\mu} - \bar{\boldsymbol{x}})(\boldsymbol{\mu} - \bar{\boldsymbol{x}})^{\mathrm{T}} + N\bar{\mathbf{R}}\right]^{-1}$
  où $\bar{\mathbf{R}} = \sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^{\mathrm{T}}$ est la matrice de covariance empirique,
- $\eta^{\text{post}} = \eta + N$,
- $\nu^{\text{post}} = \nu + N$.

<u>échantillonnage :</u> explicite

L'algorithme est composé d'une boucle de Gibbs qui emboîte une boucle contenant l'échantillonnage des paramètres attachés aux expérience. À l'intérieur de la boucle expérience à l'itération (k), nous échantillonnons pour l'expérience indicée par n les paramètres $(\boldsymbol{x}_n)^{(k)}$, $(\boldsymbol{\kappa}_n)^{(k)}$ et $(\boldsymbol{\theta}_n^{\text{inst}})^{(k)}$ indépendamment des autres expériences. Ensuite, à la fin de la boucle intérieure, nous tirons un échantillon pour le couple $(\boldsymbol{\pi})^{(k)}$.

Remarque ⟨5.5⟩ (Parallélisation) *L'échantillonnage des paramètres $\boldsymbol{x}_n$, $\boldsymbol{\kappa}_n$ et $\boldsymbol{\theta}_n^{\text{inst}}$ peut être parallélisé. En effet, l'échantillonnage pour l'expérience n est indépendant des autres expériences ce qui permet de distribuer le travail d'échantillonnage sur plusieurs instances de calcul (CPU, GPU ou ordinateurs en réseau).* MATLAB *propose pour cela des boîtes à outils (Parallel Computing Toolbox [109]), mais il existe également des solutions gratuites avec* GPUmat *[110] pour le calcul sur GPU et* Multicore *[111] pour le calcul sur plusieurs CPU et/ou ordinateurs ayant accès au même répertoire.*

Nous arrêtons l'échantillonnage après $K_0 + K$ itérations. Si on écrit $\left[\boldsymbol{\pi}, \boldsymbol{x}_{1:N}, \boldsymbol{\kappa}_{1:N}, \boldsymbol{\theta}_{1:N}^{\text{inst}}\right]$ le vecteur complet de paramètres, la marginalisation des paramètres $\boldsymbol{x}_{1:N}$, $\boldsymbol{\kappa}_{1:N}$ et $\boldsymbol{\theta}_{1:N}^{\text{inst}}$ se fait par la projection du vecteur complet sur la composante $\boldsymbol{\pi}$ pour tous les vecteurs complets échantillonnés. Autrement dit, nous omettons les échantillons $(\boldsymbol{x}_{1:N})^{(k)}$, $(\boldsymbol{\kappa}_{1:N})^{(k)}$ et $(\boldsymbol{\theta}_{1:N}^{\text{inst}})^{(k)}$ pour ne garder que $(\boldsymbol{\pi})^{(k)}$. Ainsi, nous intégrons la toutes les valeurs échantillonnées dans celles pour $\boldsymbol{\pi}$, ce qui est équivalent à la marginalisation.

À partir des échantillons marginaux pour $\boldsymbol{\pi}$, nous calculons l'EAP en moyennant les échantillons (k) tel que $k > K_0$.

L'algorithme est résumé dans Alg. 5.1, page 123.

Remarque ⟨5.6⟩ (« Rien ne se perd, … tout se transforme. ») *Nous avons porté toute notre attention à l'estimation du paramètre d'intérêt $\boldsymbol{\pi}$. Nous pouvons également, sans coût supplémentaire, récupérer les estimations des paramètres de nuisance $\boldsymbol{x}_n$, $\boldsymbol{\kappa}_n$ et $\boldsymbol{\theta}_n^{\text{inst}}$ pour les expériences $n = 1, \ldots, N$. Pour cela, il suffit de moyenner les échantillons correspondants qu'il a de toute façon fallu tirer pour permettre l'estimation des paramètres de la classe.*

5.3.5 Résultats

Les performances de l'apprentissage peuvent être quantifiées uniquement avec des données pour lesquelles on maîtrise précisément l'entrée. Quant aux données synthétiques, *primo* nous n'en disposons pas suffisamment. *Secundo*, elles ne nous permettraient pas de comparer les distributions puisqu'il s'agit d'une campagne expérimentale

de type rampe. Ainsi, la seule information que nous maîtrisions (et ce seulement à des erreurs de préparation près) est la concentration injectée. La précision étant cependant est inconnue.

Il nous reste alors des simulations où nous maîtrisons la distribution au niveau de la cohorte et ainsi les paramètres moyenne et précision. Nous sommes donc contraints dans cette section de travailler avec des données simulées, à la fois pour le Full-MS et pour le SRM.

Présentation

Étude sur données simulées en mode Full-MS — N'ayant pas de cohorte de données d'effectif assez grand, nous avons choisi d'en simuler à une classe de taille $N \in \{10, 20, 50, 100, 200, 1000\}$, en utilisant les paramètres de génération de données de la Sect. 3.6. Ceci permettra d'évaluer l'efficacité de l'apprentissage dans ce cadre. Des indications par rapport à l'effectif efficient pour une cohorte de données réelles pourront être déduites seulement après l'étude d'une telle cohorte. De plus, cela dépend aussi d'un certain nombre de facteurs comme la qualité du signal et le niveau de bruit.

Nous évaluerons trois distributions.

1. Nous noterons $\mathring{\mathrm{p}}$ la distribution vraie de la cohorte. [4]
2. La distribution $\tilde{\mathrm{p}}$ est celle issue d'un apprentissage *partiel*. Lors de cet apprentissage partiel, nous fixons trois des paramètres techniques $[\boldsymbol{\xi}_{1:N}, \boldsymbol{\tau}_{1:N}, \gamma_{1:N}]$ aux valeurs vraies dans l'algorithme. Nous réduisons ainsi la variabilité technique. Les seules sources de variabilité qui subsistent sont celle liées à la phase de digestion et celle liée à la concentration.
3. Enfin, la distribution $\hat{\mathrm{p}}$ désigne la distribution résultant de l'apprentissage global, intégrant tous les paramètres et ainsi toutes les sources de variabilités identifiées. Elle correspond à ce que nous avons présenté ci-dessus, estimant l'ensemble des paramètres.

Pour chaque individu n, nous simulons la mesure $\mathbf{Y}_n$ de la protéine *Entérotoxine A* du Staphylocoque doré (SEA) dont les propriétés ont été enumérées dans [1]. Comme il s'agit d'une seule protéine, nous ne travaillerons plus avec les vecteurs et les matrices, mais avec des scalaires. Les caractéristiques stables et non aléatoires dans notre cadre sont résumées dans la Tab. 3.1.

Les paramètres de classe sont donnés par $(\boldsymbol{m}, \boldsymbol{\Gamma}) \equiv (m, \gamma) = (50, 0.2)$. En ce qui concerne les paramètres aléatoires, ils ont été simulés de la manière suivante. La concentration du biomarqueur x_n suit une loi normale de moyenne m et de précision γ. Ne suivant qu'un peptide, sa quantité κ_n suit une loi normale de moyenne x_n et de précision γ_κ. Ce peptide subit un gain d'ionisation ξ_n qui suit une loi normale de moyenne 10 et de précision 10. Le temps de rétention τ_n est choisi uniformément dans l'intervalle $[\tau^{\mathrm{m}}, \tau^{\mathrm{M}}] = [1628.5, 1668.5]$ s. Finalement, le niveau de bruit γ_n est tiré uniformément dans l'intervalle $[0.01, 1]$. Ainsi, nous pouvons simuler une grande palette de données, certaines très perturbées par le bruit de mesure ou par d'autres processus aléatoires.

Dans la suite, nous donnons des résultats d'apprentissage moyens, c'est-à-dire calculés à partir de la description précédente et pour un N donné. Nous avons simulé 200 cohortes. Quant à l'inversion, nous choisissons des lois vaguement informatives pour le couple $(\boldsymbol{m}, \boldsymbol{\Gamma})$, les paramètres $\boldsymbol{\xi}_n$ et γ_n. Nous avons ensuite procédé à l'apprentissage de ces cohortes, dont nous avons moyenné les résultats pour obtenir une divergence moyenne, une moyenne moyenne et une précision moyenne par effectif.

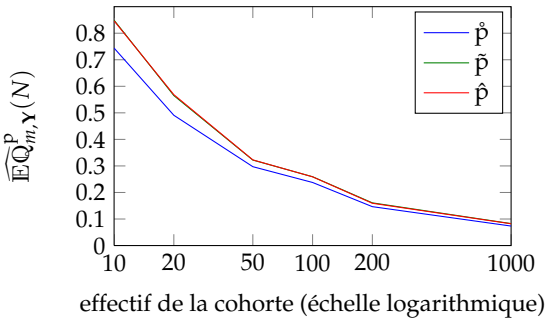
4. En biostatistique on parle de la distribution de l'échantillon. Avec $N \to \infty$ ou $N \to N_{\max}$, l'échantillon tend vers la population et $\mathring{\mathrm{p}} \to \mathrm{p}^\star$.

Tab. 5.1 Tableaux comparatifs des résultats de l'apprentissage à partir de données Full-MS pour les distribution $\mathring{p}$, $\tilde{p}$ et $\hat{p}$. Attention : dans les trois figures, les courbes vertes ($\tilde{p}$) et rouges ($\hat{p}$) sont superposées !

(a) Estimation de la moyenne, $m^\star = 50$

N	$\mathring{p}$	$\tilde{p}$	$\hat{p}$
10	50.16	50.15	50.15
20	50.07	50.08	50.08
50	49.98	49.97	49.98
100	50.01	50.00	50.00
200	49.99	49.97	49.98
1000	50.00	50.00	50.00

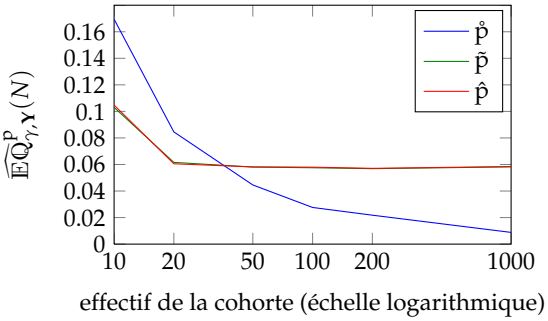
(b) Erreur par rapport à $m^\star = 50$



(c) Estimation de la precision, $\gamma^\star = 0.2$

N	$\mathring{p}$	$\tilde{p}$	$\hat{p}$
10	0.27	0.19	0.19
20	0.23	0.16	0.16
50	0.21	0.15	0.15
100	0.20	0.14	0.14
200	0.20	0.14	0.14
1000	0.20	0.14	0.14

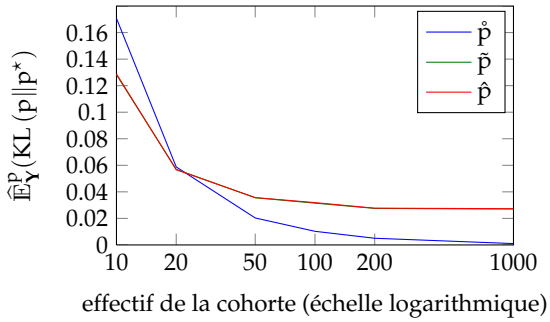
(d) Erreur par rapport à $\gamma^\star = 0.2$



(e) Divergence de Kullback-Leibler

N	$\mathrm{KL}\,(\mathring{p}\|p^\star)$	$\mathrm{KL}\,(\tilde{p}\|p^\star)$	$\mathrm{KL}\,(\hat{p}\|p^\star)$
10	0.17	0.13	0.13
20	0.06	0.06	0.06
50	0.02	0.04	0.04
100	0.01	0.03	0.03
200	0.01	0.03	0.03
1000	0.00	0.03	0.03

(f) Évolution de la divergence de Kullback-Leibler



Dans les Tab. 5.1(a) et Tab. 5.1(c), nous exposons les résultats moyens de l'estimation de la moyenne et de la précision de la classe, ainsi que les erreurs relatives, en fonction de l'effectif. La Tab. 5.1(e) présente les divergence de Kullback-Leibler moyenne entre la distribution vraie $p^\star$ et les distributions apprises. Nous représentons cette table dans la Fig. 5.1(f) où l'abscisse correspond à l'effectif en échelle logarithmique et l'ordonnée à la divergence de Kullback-Leibler associée. On voit sur les figures associées aux tables que l'erreur sur la moyenne décroît continûment quand N devient grand, et que l'écart entre distributions apprises partiellement ou globalement et la vérité de cohorte devient petit (Fig. 5.1(b)). L'erreur sur la précision se comporte différemment : alors qu'elle décroît pour la vérité de la cohorte conformément à nos attentes, elle atteint un plateau pour les méthodes d'inversion à partir de $N = 20$ (Fig. 5.1(d)). En combinant ces résultats, la divergence de Kullback-Leibler décroît vite pour la vérité de la cohorte, mais décroît lentement pour les distributions apprises par inversion.

On voit également très bien sur les courbes traçant les erreurs sur la moyenne et sur la précision et la divergence que les apprentissages partiel et global atteignent les mêmes performances, les courbes étant superposées.

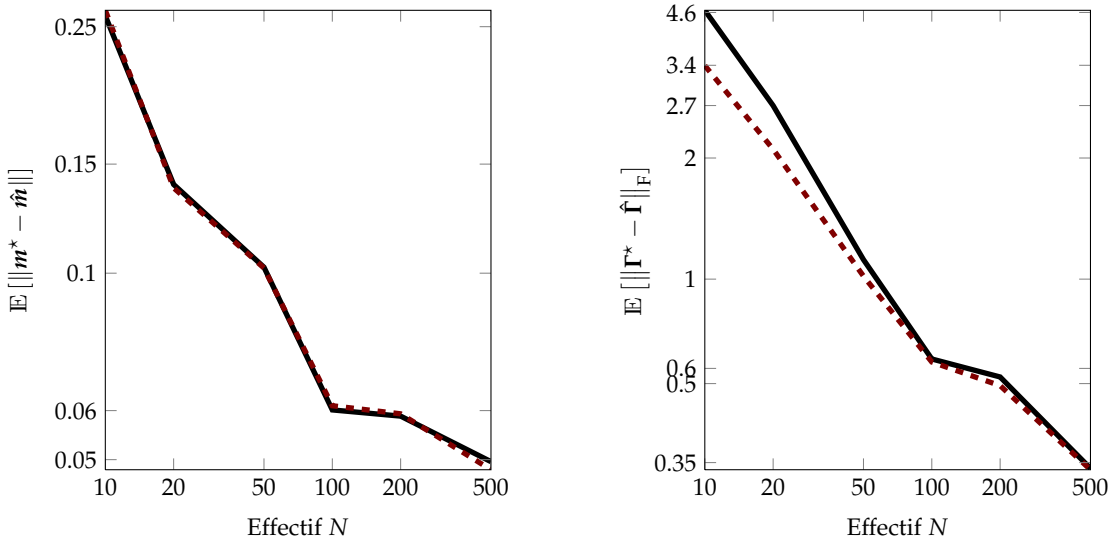
Étude sur données simulées en mode SRM — Nous considérons le cas décrit par le schéma suivant : $\boxed{P = 2} \xrightarrow{1 \quad 3} \boxed{I = 6} \xrightarrow{1 \quad 3} \boxed{L = 18}$, la distribution des protéines au niveau de la population étant normale de moyenne $\boldsymbol{m}^\star = [3, 3]^{\mathrm{T}}$ et de précision $\boldsymbol{\Gamma}^\star = \left[\begin{smallmatrix} 3 & 0.5 \\ 0.5 & 3 \end{smallmatrix}\right]$. Les autres paramètres sont modulés pour imiter les données réelles le plus fidèlement possible. À titre d'information, l'ordre de grandeur du gain de fragmentation est 10^4 avec une variabilité de 10 %, celui des inverses variances du bruit de mesures est 10^{-5}, les paramètres chromatographiques et le gain de transition sont simulés selon les distributions (uniformes, normale) correspondantes. Nous avons répété l'apprentissage vingt fois pour les cohortes de tailles différentes ($N \in \{10, 20, 50, 100, 200, 500\}$).

Nous avons montré dans la section précédente et dans les illustrations associées que l'inversion partielle et l'inversion globale atteignent les mêmes performances. C'est la raison pour laquelle dans la présente section, nous ne considérons que les résultats de l'inversion globale et les comparons à la vérité de la cohorte.

La Fig. 5.3(a) montre l'erreur d'estimation de la moyenne et dans la Sous-Fig. (b) l'erreur d'estimation de la précision en fonction de l'effectif de la cohorte, toujours par rapport à la vérité de la population indiquée ci-dessus, en trait plein. Le trait rouge pointillé correspond à l'erreur de la vérité de cohorte par rapport à la population. On voit que les courbes sont très proches l'une de l'autre, se superposant quasi totalement pour l'erreur sur la moyenne dès un petit effectif. Les courbes se confondent partiellement à partir de $N = 100$ pour l'erreur sur la précision, cette dernière étant surestimée pour de petits effectifs.

La différence des reconstructions des distributions des classes est visible dans Fig. 5.4 où nous traçons la vérité de la cohorte en noir, la reconstruction de la distribution apprise avec un effectif de $N = 10$ individus en rouge, et avec $N = 500$ individus en vert. On voit clairement que l'apprentissage avec un faible nombre d'individu approche certes la moyenne de la population d'une manière relativement faible, mais la matrice de précision est surestimée, les interactions sont mal prises en compte ($\hat{\boldsymbol{m}}_{10} = [3.03, 2.92]^{\mathrm{T}}$, $\hat{\boldsymbol{\Gamma}}_{10} = \left[\begin{smallmatrix} 5.43 & -0.08 \\ -0.08 & 6.15 \end{smallmatrix}\right]$). La reconstruction avec un effectif de $N = 500$ individus superpose la vérité de la cohorte. Elle est proche de la distribution de la population, à la fois en terme de moyenne mais aussi en terme de précision, *i.e.* la cohorte représente bien la population. Les covariances sont également bien estimées. Les valeurs estimées sont $\hat{\boldsymbol{m}}_{500} = [3.01, 2.99]^{\mathrm{T}}$, $\hat{\boldsymbol{\Gamma}}_{500} = \left[\begin{smallmatrix} 3.03 & 0.60 \\ 0.60 & 3.07 \end{smallmatrix}\right]$.

Fig. 5.3 Évaluation de l'apprentissage sur des données simulées SRM, axes logarithmiques. Les traits pleins correspondent à l'évolution de l'erreur d'estimation avec la méthode d'apprentissage proposée, les traits pointillés à l'évolution de l'erreur de la vérité de la cohorte par rapport à la vérité de la population.



(a) Moyenne (b) Précision

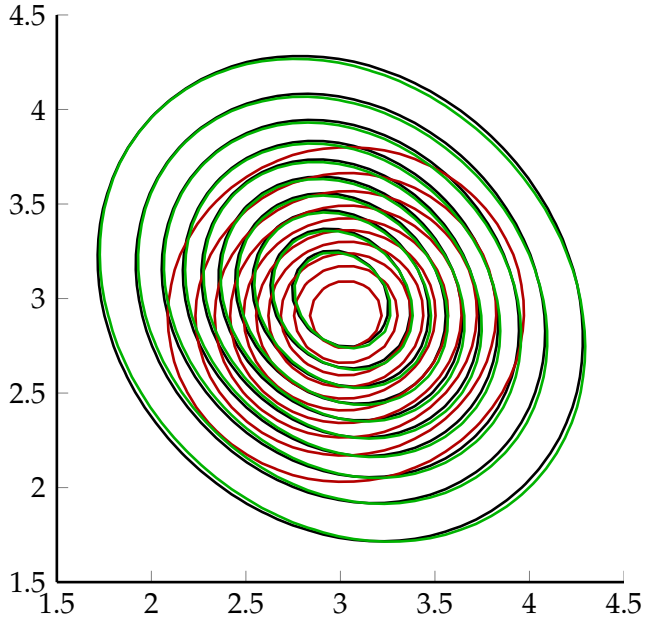
Discussion

Au vu des résultats, nous pouvons constater plusieurs choses. Quelle que soit la provenance des données, plus on utilise d'individus, meilleure est l'estimation. Ceci n'est bien évidemment pas une surprise, mais nous réconforte puisque c'est une preuve que notre méthode fonctionne raisonnablement. En effet, même si la moyenne est bien estimée quel que soit l'effectif, c'est surtout la précision qui est sensible au nombre d'individus de la classe. On voit que l'estimation de la précision se stabilise à partir d'un certain nombre d'individus, l'échantillon de la population n'étant pas assez représentatif pour de petits effectifs.

Nous séparons ensuite les données Full-MS des données SRM pour mieux commenter les résultats.

Full-MS — La décroissance de la divergence de Kullback-Leibler est presque exponentielle. Alors que la courbe dans Fig. 5.1(f) pour les données Full-MS décroît vite dans l'intervalle $[10, 50]$, elle ne le fait que lentement après à partir de 50 individus. La courbe bleue représente la divergence entre la distribution vraie et celle de la cohorte, les courbes verte et rouge entre la distribution vraie et celle apprises par l'inversion partielle et globale respectivement. On s'aperçoit que la courbe bleue chute de plus haut et plus brusquement. En plus, elle a une intersection avec les deux autres vers $N = 20$. Comme elle représente la distribution vraie à l'intérieur de la cohorte, les distributions apprises apparaissent comme des versions « régularisées », vue la pente moins importante. On voit également que la divergence pour les distributions apprises est quasiment constante à partir de $N = 100$, alors qu'elle ne cesse de chuter pour la distribution de la cohorte qui approche la distribution vraie plus N est grand. Comme nous intégrons dans l'inversion les variabilités technique et biologique, la divergence entre la distribution vraie et la distribution apprise par inversion converge vers une valeur positive non nulle, ce qui est dû à l'inclusion de l'incertitude pour chaque paramètre.

Fig. 5.4 Reconstruction des distributions des classes apprises (lignes de niveau) ; noir : distribution au niveau de la population ; rouge : distribution apprise avec $N = 10$; vert : distribution apprise avec $N = 500$



Notons aussi que les distributions apprises sont plus larges que les distributions des cohortes, comme nous pouvons le constater dans la Tab. 5.1(c) et dans la Fig. 5.1(d). Chaque mesure a son incertitude qui est ensuite transmise aux paramètres à l'aide desquels nous estimons les paramètres de classe. Le fait que les distributions soient plus larges est donc fortement lié aux variabilités technique (notamment au bruit sur l'ensemble des $N_{\mathscr{S}} \cdot N_{\mathscr{C}}$ points de mesure, ce nombre pouvant être très grand) et biologique et leur intégration dans l'estimation, comme schématisé sur la Fig. 5.5.

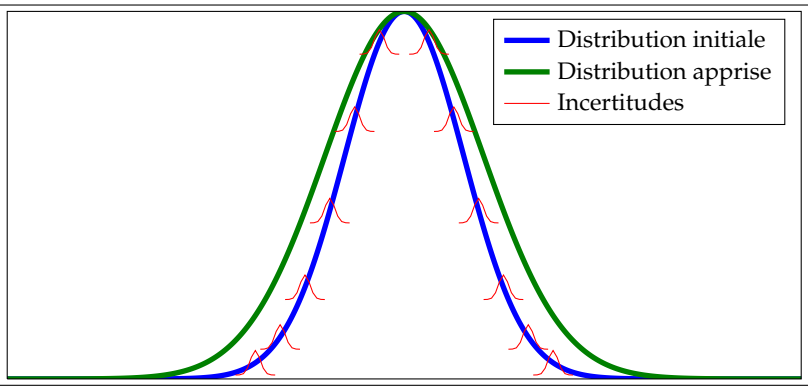
SRM — Nous avons remarqué que, pour les données LC-Full-MS, nous arrivons à un plateau au niveau de l'estimation des précisions. Il se trouve que nous ne le rencontrons pas pour les données LC-SRM, voir Fig. 5.3. Nous avons en effet simulé un bruit de mesure plus modéré. La valeur du paramètre du bruit est propagée d'abord dans la loi *a posteriori* conditionnelle de la quantité peptidique, et à travers cette dernière dans la loi *a posteriori* conditionnelle de la concentration protéique, et finalement dans la loi *a posteriori* conditionnelle des paramètres de la classe. Cette propagation explique l'existence ou l'absence d'un plateau au niveau de la précision.

On peut s'apercevoir, comme pour les données LC-Full-MS, que l'estimation avec un effectif de cent individus semble suffisamment proche de la distribution au niveau de la population. Les effectifs plus élevés n'améliorent que légèrement les résultats. Ceci peut notamment être un critère à prendre en compte lors du choix de la taille de la cohorte et caractériser ainsi un compromis « temps d'apprentissage/effectif d'apprentissage/degré de précision de l'apprentissage ».

5.4 Classer

Dans la section précédente, nous avons vu comment le classifieur est entraîné à partir de cohortes étiquetées. Voyons maintenant comment cette connaissance nous est utile dans l'application du classifieur à un nouvel individu à classer. En effet, les paramètres

Fig. 5.5 Schématisation de l'incertitude due à la variabilité technique (représentée par les courbes rouges), la distribution initiale (bleu) et la distribution estimée (vert). La distribution estimée est plus large que la distribution initiale.



appris $\hat{\boldsymbol{\pi}}^{1:\mathsf{M}}$ précédemment sont à la base du classement puisqu'il s'agit des hyperparamètres de la distribution de la concentration des protéines, ces hyperparamètres étant déterminés par la classe en question. Notons donc, vu le lien imposé entre classe et paramètres, qu'il est équivalent d'écrire $\mathrm{p}(\boldsymbol{x} \mid \hat{\boldsymbol{\pi}}^{\mathsf{m}})$, $\mathrm{p}(\boldsymbol{x} \mid \mathrm{c}^{\mathsf{m}})$ ou $\mathrm{p}(\boldsymbol{x} \mid \mathrm{c}^{\mathsf{m}}, \hat{\boldsymbol{\pi}}^{\mathsf{m}})$. Dans ce qui suit, nous allons omettre la marque "chapeau" sur la variable des paramètres des classes tout en gardant à l'esprit qu'il s'agit d'une estimation.

Voici la situation pour le problème de classement :

- les paramètres des classes $\boldsymbol{\pi}^{1:\mathsf{M}}$ et la proportion des classes regroupé dans le vecteur $[p_1, \ldots, p_{\mathsf{M}}] = \boldsymbol{p}$ sont connus ;
- l'observation $\mathbf{Y}$ provient d'un individu de classe indéterminée.

À quelle classe appartient cet individu ? Dans un contexte clinique, la question qui est posée se traduit par : quel diagnostic devons-nous faire pour l'individu compte tenu de son échantillon biologique ? Quel est son statut clinique ? Pour répondre à cette question, nous allons construire le classifieur $\psi^{\mathrm{c}} : \Omega_{\mathbf{Y}} \to \mathcal{C}$ qui associe à une observation $\mathbf{Y}$ une estimation de classe $\psi^{\mathrm{c}}(\mathbf{Y}) = \hat{\mathrm{c}}$.

5.4.1 Modèle direct et loi jointe

Considérons la situation telle qu'elle est décrite dans la Fig. 5.2(b). Suivant ce schéma et reprenant les hypothèses d'indépendance énoncées précédemment, nous pouvons immédiatement déduire pour la loi jointe

$$\mathrm{p}(C, \boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}}, \mathbf{Y}) = \underbrace{\Pr(C)}_{\text{classe}} \underbrace{\mathrm{p}(\boldsymbol{x} \mid C)\, \mathrm{p}(\boldsymbol{\kappa} \mid \boldsymbol{x})\, \mathrm{p}(\boldsymbol{\theta}^{\mathrm{inst}})\, \mathrm{p}(\mathbf{Y} \mid \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}})}_{\text{expérience}}. \tag{5.15}$$

On retrouve à nouveau, comme nous l'avons déjà fait remarquer, que le groupe de facteurs « expérience » est identique (à un conditionnement près) à la loi jointe de la quantification. On reprend les lois proposées précédemment pour ces facteurs.

Nous choisissons pour la distribution pour la classe une distribution catégorielle, identifiée par le vecteur $\Pr(C) = \big[\Pr(C = \mathrm{c}^1), \ldots, \Pr(C = \mathrm{c}^{\mathsf{M}})\big] = [p_1, \ldots, p_{\mathsf{M}}] = \boldsymbol{p}$. Chaque probabilité de classe est associée à sa proportion dans la population.

Remarque ⟨**5.7**⟩ *Dans cette thèse, nous n'avons pas analysé l'estimation de cette loi catégorielle. Cela étant dit, elle peut s'apprendre en même temps que les paramètres des classes dans l'étape d'apprentissage, ayant à disposition une cohorte suffisamment grande. Pour cela, une étude [90] a été mené dans un cas binaire. Dans ce cas, la distribution catégorielle qui devient une distribution de Bernoulli conjugue la loi Bêta. C'est cette dernière que l'on choisit comme distribution* a priori

sur la proportion p d'une classe. La loi a posteriori *pour p fait intervenir dans ses arguments les estimateurs empiriques de la proportion, i.e. nombre d'échantillons de la classe donnée par rapport au nombre d'échantillons total.*

5.4.2 Expression de l'estimateur

Nous cherchons à estimer la classe associée à une observation $\mathbf{Y}$ donnée. Le classifieur étant construit sur un raisonnement bayésien, nous devons exprimer la loi *a posteriori* $\Pr(C \mid \mathbf{Y})$ ce qui appelle, comme nous l'avons déjà fait dans la section précédente, la marginalisation des paramètres de nuisance :

$$\Pr(C \mid \mathbf{Y}) = \int_{\Omega_{\boldsymbol{x},\boldsymbol{\kappa},\boldsymbol{\theta}^{\text{inst}}}} \mathrm{p}(C, \boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\text{inst}} \mid \mathbf{Y})\, \mathrm{d}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\text{inst}}). \tag{5.16}$$

La marginalisation permet de considérer toutes les valeurs conjointement possibles *a posteriori* pour les paramètres de nuisance. Nous explorons ainsi tout l'espace de probabilité $\Omega_{\boldsymbol{x},\boldsymbol{\kappa},\boldsymbol{\theta}^{\text{inst}}}$ des concentrations protéiques, quantités peptidiques et paramètres instruments — contrairement aux approches séquentielles qui utilisent des estimations ponctuelles intermédiaires.

À partir de l'expression de cette loi – qui est toujours une distribution catégorielle – l'observation $\mathbf{Y}$ peut être classée dans une des classes considérées $\mathrm{c}^1, \ldots, \mathrm{c}^{\mathsf{M}}$. Pour ce faire, plusieurs choix sont possibles qui sont du type « test d'hypothèse, choix de modèle » que nous avons introduits dans la Sect. 2.6. Nous allons notamment nous concentrer sur le classifieur défini à partir d'une fonction de coût 0–1. Nous rappelons qu'il s'agit de l'estimateur Maximum A Posteriori pour une distribution discrète qui est donnée par l'Éqn. (2.16) :

$$\psi^{\mathrm{c}}(\mathbf{Y}) = \operatorname*{argmax}_{\mathsf{m}=1,\ldots,\mathsf{M}} \Pr(C = \mathrm{c}^{\mathsf{m}} \mid \mathbf{Y}) \tag{2.16}$$

Remarque ⟨5.8⟩ (Non-décision) *Que faire dans un cas où la décision d'appartenance à une classe ne peut pas être prise ou la prise d'une décision est trop incertaine ? En effet, il y a des cas où l'apport des données n'est pas assez concluant (signaux trop bruités, de mauvaise qualité, ...) pour répondre à la question de classification avec les classes proposées dans la vérité terrain $\mathcal{C}$ qui est l'ensemble du départ du modèle hiérarchique. Pour répondre à cette question, considérons les deux propositions suivantes.*

1. *Nous étendons l'ensemble d'arrivée du classifieur $\mathcal{C}$ en ajoutant une classe artificielle* ND, $\widetilde{\mathcal{C}} = \mathcal{C} \cup \{\mathrm{ND}\}$. *La loi* a priori *normale de la concentration des biomarqueurs associée à cette classe est choisie très faiblement informative. En dehors des « zone à forte probabilité » des classes de la vérité terrain $\mathcal{C}$, cette densité est plus grande car elle s'estompe plus lentement. Autrement dit, là où on n'est pas sûr d'observer un processus issu d'une classe « connue » compte tenu des distributions* a priori, *un processus* ND *est plus probable. Cette approche d'étendre l'ensemble se combine parfaitement avec ce qui précède.*
2. *On définit un intervalle de non décision $]r, a[$ pour le facteur de Bayes [2, Note 5.7.4]. Par contre, ceci implique que l'ensemble d'arrivée du classifieur n'est pas, comme ci-dessus, égal à l'ensemble de départ $\mathcal{C}$, mais à son extension $\mathcal{C} \cup \{\mathrm{ND}\}$. Ainsi, la règle de classification devient dans l'exemple binaire*

$$\hat{\mathrm{c}} = \begin{cases} 0 & \textit{si } \mathrm{BF}_{01} \geq a, \\ 1 & \textit{si } \mathrm{BF}_{01} \leq r, \\ \mathrm{ND} & \textit{si } r < \mathrm{BF}_{01} < a. \end{cases}$$

Que ce soit en utilisant la première ou la deuxième méthode, la réponse ND *indique que nous avons besoin d'une nouvelle acquisition des données (qui auront peut-être de meilleures conditions, plus de minutie dans la préparation, moins de défauts instrumentaux, une meilleure réalisation des pré-traitements, etc.) ou d'informations* a priori *plus exactes, plus précises.*

Dans le cas où l'on souhaite estimer également les paramètres de nuisance, le quantifieur ψ^{q} attaché au coût quadratique est l'espérance de la loi *a posteriori* conditionnée par la sortie du classifieur ψ^{c} pour éviter un mélange de modèle :

$$\left[\hat{\boldsymbol{x}}, \hat{\boldsymbol{\kappa}}, \hat{\theta}^{\mathrm{inst}}\right] = \psi^{\mathrm{q}}(\mathbf{Y}) = \int\limits_{\Omega_{\boldsymbol{x},\boldsymbol{\kappa},\boldsymbol{\theta}^{\mathrm{inst}}}} \left[\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}}\right] \mathrm{p}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}} \mid \mathbf{Y}, \psi^{\mathrm{c}}(\mathbf{Y}))\, \mathrm{d}(\boldsymbol{x}, \boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}}). \tag{5.17}$$

La marginalisation dans Éqn. (5.16) et le calcul du premier moment de la loi *a posteriori* conditionnelle dans Éqn. (5.17) ne sont pas possibles analytiquement dans ce cas dû à la complexité du problème posé. Nous proposons cependant d'utiliser les méthodes de Monte-Carlo pour les calculs des intégrales, attachées à une approximation des lois par des chaînes de Markov, comme nous l'avons déjà fait pour résoudre les problèmes d'intégrations précédents.

5.4.3 Mise en œuvre de l'inversion

Dans cette section, nous présentons la mise en œuvre algorithmique des calculs introduits précédemment. Notons premièrement que nous adoptons une structure de Gibbs (Sect. 2.9.1) pour passer du problème d'échantillonnage global à des sous-problèmes d'échantillonnage plus faciles. Ensuite, la structure hiérarchique simplifie encore les expressions : grâce à la propagation de l'information à la hiérarchie, les paramètres intervenant dans la loi *a posteriori* conditionnelle sont réduits à ceux interagissant directement avec le paramètre en question, *i.e.* les hyperparamètres (niveau hiérarchique supérieur), les paramètres voisins (niveau hiérarchique égal) nécessaires pour donner le résultat, et le résultat même (niveau hiérarchique inférieur) (voir Sect. 2.7). Les indépendances hiérarchiques, introduites par la physique et traduites dans le modèle direct, permettent donc un calcul plus simple et plus efficace.

Lois a posteriori conditionnelles pour l'étiquette de la classe

Compte tenu de la factorisation de la loi jointe du problème du classement d'une observation en facteurs *classe* et *expérience*, nous pouvons reposer les développements des lois *a posteriori* conditionnelles pour les paramètres instruments et la quantité peptidique sur ce que nous avons introduit dans le Ch. 4 (Inversion-Quantification). Même si la distribution pour la concentration protéique ne change pas de forme de loi en soi si elle est conditionnée par une classe, il est important de mentionner que la distribution marginale pour $\boldsymbol{x}$ est un mélange de lois normales :

$$\mathrm{p}(\boldsymbol{x}) = \sum_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{p}(\boldsymbol{x}, C = \mathrm{c}^{\mathrm{m}}) = \sum_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{p}(\boldsymbol{x} \mid C = \mathrm{c}^{\mathrm{m}})\, \Pr(C = \mathrm{c}^{\mathrm{m}}).$$

La distribution catégorielle étant discrète, le calcul de la distribution *a posteriori* peut se faire analytiquement et est également une distribution discrète. Compte tenu de ces faits, nous résumons les lois *a posteriori* conditionnelles pour la concentration et la classe. La déduction des lois *a posteriori* des autres paramètres est identique à celle présentée dans la Sect. 4.3.1 (Inversion-Quantification, Lois a posteriori conditionnelles), page 72.

(1) concentration des protéines cibles $\boldsymbol{x}$

<u>a priori :</u> $\mathrm{p}(\boldsymbol{x} \mid \mathrm{c}^{\mathsf{m}}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_{\mathsf{m}}, \boldsymbol{\Gamma}_{\mathsf{m}})$
<u>vraisemblance associée :</u> $\mathrm{p}(\boldsymbol{\kappa} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{\kappa}; \mathbf{D}\boldsymbol{x}, \boldsymbol{\Gamma}_{\boldsymbol{\kappa}})$
<u>conjugaison :</u> oui car $\boldsymbol{x}$ est linéaire dans les arguments des deux lois normales
<u>a posteriori :</u> $\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\kappa}, \mathrm{c}^{\mathsf{m}}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_{\boldsymbol{x}}^{\mathrm{post}}, \boldsymbol{\Gamma}_{\boldsymbol{x}}^{\mathrm{post}})$
- $\boldsymbol{m}_{\boldsymbol{x}}^{\mathrm{post}} = (\boldsymbol{\Gamma}^{\mathrm{post}})^{-1}(\boldsymbol{\Gamma}_{\mathsf{m}}\boldsymbol{m}_{\mathsf{m}} + \mathbf{D}^{\mathrm{T}}\boldsymbol{\Gamma}_{\boldsymbol{\kappa}}\boldsymbol{\kappa}_n)$,
- $\boldsymbol{\Gamma}_{\boldsymbol{x}}^{\mathrm{post}} = \boldsymbol{\Gamma}_{\mathsf{m}} + \mathbf{D}^{\mathrm{T}}\boldsymbol{\Gamma}_{\boldsymbol{\kappa}}\mathbf{D}$;

<u>échantillonnage :</u> explicite

(2) classe c

<u>a priori :</u> $\Pr(C = \mathrm{c}^{\mathsf{m}}) = p_{\mathsf{m}}$
<u>vraisemblance associée :</u> $\mathrm{p}(\boldsymbol{x} \mid C = \mathrm{c}^{\mathsf{m}}) = \mathcal{N}(\boldsymbol{x}_{\mathsf{m}}; \boldsymbol{m}_{\mathsf{m}}, \boldsymbol{\Gamma}_{\mathsf{m}})$
<u>conjugaison :</u> oui
<u>a posteriori :</u> $\mathrm{p}(C = \mathrm{c}^{\mathsf{m}} \mid \boldsymbol{x}) = p_{\mathsf{m}}^{\mathrm{post}}$
- $p_{\mathsf{m}}^{\mathrm{post}} = \dfrac{p_{\mathsf{m}}|\boldsymbol{\Gamma}_{\mathsf{m}}|^{P/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_{\mathsf{m}})^{\mathrm{T}}\boldsymbol{\Gamma}_{\mathsf{m}}(\boldsymbol{x}-\boldsymbol{m}_{\mathsf{m}})\right)}{\sum_{n=1}^{\mathsf{M}} p_n|\boldsymbol{\Gamma}_n|^{P/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_n)^{\mathrm{T}}\boldsymbol{\Gamma}_n(\boldsymbol{x}-\boldsymbol{m}_n)\right)}$

<u>échantillonnage :</u> explicite

Calcul de la marginalisation des paramètres de nuisance

Nous présentons deux moyens (Chib, échantillonnage explicite) de calculer la loi marginale parmi tant d'autres (*Bridge sampling*, échantillonnage d'importance, ... [2, sect. 7.3]). La première consiste à lancer autant de chaînes d'échantillonnage que l'on a de classes, en fixant pour chaque chaîne la classe et ainsi les lois *a priori*. Nous récupérons ainsi une vraisemblance marginale qui est liée par la règle de Bayes à la probabilité a posteriori marginale que nous cherchons à exprimer. Dans la deuxième méthode, nous échantillonnons directement la classe et calculons ensuite la probabilité a posteriori à l'aide des échantillons. Notons premièrement que les deux approches relèvent des méthodes de « choix de modèles ».

Marginalisation par Moyenne Harmonique

Soit $\mathrm{c} \in \mathcal{C}$ une classe donnée. Nous cherchons à exprimer la probabilité marginale $\Pr(\mathrm{c} \mid \mathbf{Y}) \propto \Pr(\mathrm{c}) \cdot \mathrm{p}(\mathbf{Y} \mid \mathrm{c})$. Comme $\Pr(\mathrm{c})$ est connue et que la somme des probabilités pour c est égale à 1, le facteur prépondérant du produit est l'**évidence** $\mathrm{p}(\mathbf{Y} \mid \mathrm{c})$, *i.e.* une sorte de vraisemblance par rapport à la classe.[5]

Nous pouvons approximer sa valeur par

$$\mathrm{p}(\mathbf{Y} \mid \mathrm{c}) = \left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\mathrm{p}\left(\mathbf{Y} \mid \boldsymbol{\kappa}^{(k)}, (\boldsymbol{\theta}^{\mathrm{inst}})^{(k)}\right)}\right]^{-1}, \quad [\boldsymbol{\kappa}^{(k)}, (\boldsymbol{\theta}^{\mathrm{inst}})^{(k)}] \sim \mathrm{p}(\boldsymbol{\kappa}, \boldsymbol{\theta}^{\mathrm{inst}} \mid \mathbf{Y}). \tag{5.18}$$

Plus le nombre d'échantillons *a posteriori* est grand, meilleure est l'approximation. Pour en savoir plus de cette méthode de marginalisation, voir Ann. C.1.

Nous lançons donc pour chacune des classes $\mathrm{c} \in \mathcal{C}$ une chaîne MCMC et tirons des échantillons sous la loi *a posteriori* des paramètres. À la fin de l'échantillonnage, nous calculons l'évidence de la classe $\mathrm{p}(\mathbf{Y} \mid \mathrm{c})$. Enfin, nous pouvons récupérer la probabilité *a posteriori* $\Pr(\mathrm{c} \mid \mathbf{Y})$ de la classe c.

5. L'évidence est un terme plus connue dans les approches *Choix de modèles*.

Marginalisation par échantillonnage explicite de la classe

Cette méthode propose d'échantillonner la totalité des paramètres, sans imposer une valeur fixe pour le paramètre discret c ce qui permet de ne lancer qu'une chaîne d'échantillonnage stochastique. Nous procédons à un échantillonnage joint de paramètres discrets et paramètres continus. À chaque itération, nous tirons une valeur pour c dans son ensemble $\mathcal{C}$, conditionnellement au dernier échantillon pour la concentration des protéines $\boldsymbol{x}$.

À l'arrêt de l'algorithme après $K_0 + K$ itérations, nous disposons de K échantillons distribués sous la loi jointe *a posteriori* où K_0 désigne le temps de chauffe. Afin de calculer la probabilité *a posteriori* de la classe, il suffit de calculer l'histogramme marginal pour la variable c à partir des K échantillons joints : l'opérateur de projection de l'échantillon sous forme vectorielle sur sa coordonnée c pour les échantillons correspond à la marginalisation des paramètres de nuisance dans la densité de probabilité jointe. Ainsi, nous obtenons $\Pr(C = \mathsf{c}^{\mathsf{m}} \mid \mathbf{Y})$ pour $\mathsf{m} = 1, \ldots, \mathsf{M}$.

Remarque ⟨5.9⟩ (MCMC à Saut Reversible) *Dans un cadre « sélection de modèle », ce que nous présentons dans cette sous-section peut être vu comme une méthode modifiée d'un MCMC à saut réversible (Reversible Jump MCMC, voir [2, Sect. 7.3.4], [3, Sect. 13.1], [91, Sect. 6.7], [21, Sect. 11.2], ou l'article précurseur [92]) à* M *modèles*

$$\mathcal{M}_{\mathsf{m}} : \quad \boldsymbol{x} \mid \mathsf{c}_{\mathsf{m}} \sim \mathcal{N}(\boldsymbol{m}_{\mathsf{m}}, \boldsymbol{\Gamma}_{\mathsf{m}}),$$

pour m *allant de* 1 *à* M. *Ici, les modèles ont la même dimension et ne diffèrent que dans les paramètres des classes. La probabilité d'acceptation qui détermine si on passe d'un modèle à l'autre dépend essentiellement du rapport des probabilités* a posteriori.

Les approches sont étroitement liées, même si elles ne sont pas exactement identiques. Le lecteur intéressé trouvera l'expression de la probabilité d'acceptation dans un cas binaire dans l'Ann. C.2.

Remarque ⟨5.10⟩ (Décision aléatoire, échantillonnage de la classe) *L'échantillonnage de la classe* c *peut aussi être vu comme une prise de décision aléatoire [22, Sect. 3.1]. Cependant, notre classifieur ne contient pas d'« instruction de décision » non aléatoire pour les valeurs de la variable décidante (comme dans* Example *3.5 de l'ouvrage cité), mais uniquement des instructions aléatoires, quelle que soit la valeur.*

5.4.4 Calcul de l'estimation des paramètres de nuisance

Une fois la classe fixée à $C = \hat{\mathsf{c}} = \psi^{\mathsf{c}}(\mathbf{Y})$, nous calculons l'intégrale de l'Éqn. (5.17). Dans le cas de la marginalisation Chib, nous utilisons pour cela tous les échantillons de la chaîne d'échantillonnage qui correspond à la classe estimée. Si on marginalise par échantillonnage explicite, nous réutilisons les échantillons des itérations (k) de la chaîne d'échantillonnage tels que la classe échantillonné soit égale à la classe estimée, $\mathsf{c}^{(k)} = \hat{\mathsf{c}}$.

Soit $\mathcal{K}$ l'ensemble des indices k tel que

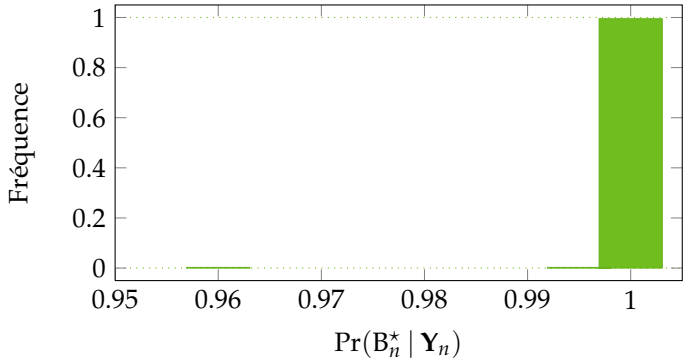
- $\mathsf{c}^{(k)} = \hat{\mathsf{c}}$ et
- $k > K_0$.

Alors la moyenne *a posteriori* conditionnelle des paramètres de nuisance s'écrit

$$\left[\hat{\boldsymbol{x}}, \hat{\boldsymbol{\kappa}}, \hat{\boldsymbol{\theta}}^{\text{inst}}\right] = \frac{1}{\text{card}(\mathcal{K})} \sum_{k \in \mathcal{K}} \left[\boldsymbol{x}^{(k)}, \boldsymbol{\kappa}^{(k)} (\boldsymbol{\theta}^{\text{inst}})^{(k)}\right]. \tag{5.19}$$

Les moments d'ordre p s'obtiennent de la même manière en moyennant les échantillons à la puissance p sous l'ensemble $\mathcal{K}$.

Fig. 5.6 Histogramme des probabilités $\Pr(\mathrm{c}_n^\star \mid \mathbf{Y}_n)$ pour $n = 1, \ldots, 400$.



5.4.5 Résultats

Dans cette section, nous évaluons l'apport de l'Inversion-Classification sur des cohortes simulées (Full-MS et SRM) et sur une cohorte clinique (SRM). Les données simulées ont été générées selon les conditions exposées dans la Sect. 3.6. La cohorte clinique correspond à celle du cancer colorectal collectée par bioMérieux en 2010 que nous avons utilisée dans la Sect. 4.5.3.

La première évalue les performances de la classification quand la distance entre les classes diminue. Pour cela, nous rapprochons et superposons les classes ce qui est un cadre réaliste pour la plupart des problèmes de classification (comme on le verra en analysant les données cliniques).

Ensuite, sur des données simulées SRM, nous caractérisons l'Inversion-Classification par rapport à la robustesse au bruit de mesure, comparée aux méthodes de classification de l'état de l'art exposées dans ce chapitre.

Finalement, nous proposons d'évaluer les performances du classifieur sur des données réelles où nous pouvons observer à la fois une superposition des distributions et des niveaux de bruit de mesure variants.

Présentation : données simulées Full-MS

Dans la cohorte se trouvent deux classes, une classe S d'individus sains et une classe M d'individus malades, chaque classe comportant 200 individus. Pour chaque individu n, la concentration du biomarqueur est déterminée par la réalisation d'un tirage aléatoire sous la loi

$$\mathrm{p}(x_n \mid \mathrm{c} = \mathrm{c}_n^\star) = \begin{cases} \mathcal{N}(x_n; m_\mathrm{S}^\star, \gamma_\mathrm{S}^\star) & \text{si } \mathrm{c}_n = \mathrm{S}, \\ \mathcal{N}(x_n; m_\mathrm{M}^\star, \gamma_\mathrm{M}^\star) & \text{sinon.} \end{cases} \tag{5.20}$$

Nous posons $m_\mathrm{S}^\star = 50$, $\gamma_\mathrm{S}^\star = \gamma_\mathrm{M}^\star = 0.2$, et débutons avec $m_\mathrm{M}^\star = 70$. Les données $\mathbf{Y}_n$ sont accompagnées de l'état initial $\mathrm{c}_n^\star$ uniquement pour l'évaluation.

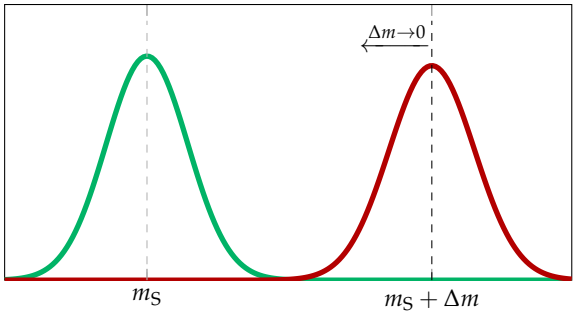
Pour la classification, nous apprenons auparavant les distributions *a priori* des classes à l'aide de l'apprentissage exposé dans la Sect. 5.3. Les résultats de cette étape où nous choisissons d'utiliser des cohortes à 100 individus d'apprentissage sont les suivants : les paramètres de la classe S sont $(m_\mathrm{S}, \gamma_\mathrm{S}) = (50.04, 0.12)$, ceux de la classe M $(m_\mathrm{M}, \gamma_\mathrm{M}) = (70.13, 0.11)$.

À l'aide de la cohorte test, nous calculons pour chaque individu $n = 1, \ldots, N = 2 \cdot 200$ la probabilité de l'état vrai de l'échantillon sachant les données, $\Pr(\mathrm{c}_n^\star \mid \mathbf{Y}_n)$. Nous représentons l'histogramme de ces résultats dans la Fig. 5.6. En utilisant le coût 0–1, tous les échantillons sont classés correctement et le risque est nul, $\mathrm{R}(\psi^\mathrm{c}) = 0$. En effet, la

Tab. 5.2 Risque du classifieur en rapprochant les moyennes des classes, *i.e.* $m_{\mathrm{M}} \to m_{\mathrm{S}} = 50.04$.

m_{M}	70.13	61.13	59.13	57.13	54.13	52.13	50.04
KL	22.4	6.83	4.60	2.79	0.931	0.244	0
$\mathrm{R}(\psi^{\mathrm{c}})$	0.00	0.00	0.04	0.09	0.20	0.34	0.50

Fig. 5.7 Illustration du rapprochement des classes par les distributions caractéristiques.



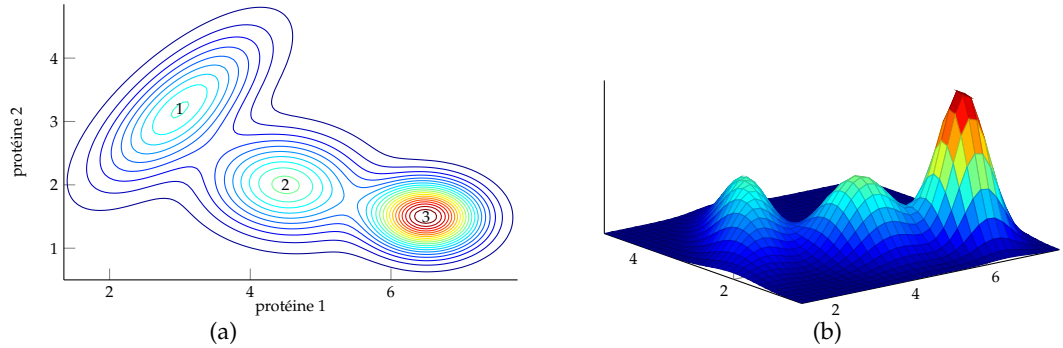
classe estimée pour tous les échantillons est la classe vraie, et pour presque tous avec une probabilité *a posteriori* de la classe de 100 %. Ainsi, la certitude que nous pouvons avoir dans l'estimation $\hat{\mathrm{c}}_n$ est presque totale.

La Tab. 5.2 montre les résultats en rapprochant les distributions des classes. Pour cela, la moyenne de la classe « malade » tend vers la moyenne de la classe « sain » (par pas de 1), la précision restant constante (mise à part la dernière étude où nous proposons l'égalité des distributions des classes). Plus les classes se rapprochent l'une de l'autre, plus les distributions se superposent, et plus les performances de la classification sont dégradées.

Discussion : données simulées Full-MS

La configuration des classes initiales ($m_{\mathrm{S}} = 50$, $m_{\mathrm{M}} = 70$, $\gamma = 0.2$) favorise naturellement la bonne estimation des classes puisque les distributions caractéristiques sont très éloignées. En effet, la superposition des distributions est infinitésimale, d'où l'étude quant à la classification quand $m_{\mathrm{M}} \to m_{\mathrm{S}}$. Évidemment, plus les classes se rapprochent l'une de l'autre, plus les distributions se superposent, et plus les performances de la classification sont dégradées ce qui est un résultat attendu. Nous pouvons pourtant constater dans la Tab. 5.2 que les performances restent raisonnables. Effectivement, le risque moyen — qui peut être interprété comme FDR (*False Discovery Rate*) si on teste les hypothèses « estimation = vérité » contre « estimation $\neq$ vérité », récupérant ainsi la fréquence de mauvaise détection — croît lentement. Pour une divergence de Kullback-Leibler supérieure à 7, le risque moyen est toujours nul. Pour des valeurs typiques du FDR, *i.e.* 5 % et 10 %, la méthode présentée dans ce chapitre est toujours efficace et satisfaisante puisque les distributions sont très proches et se superposent d'une manière conséquente. En effet, les divergences de Kullback-Leibler correspondantes valent environ 4.60 et 2.79 respectivement. Évidemment, quand les distributions sont égales, l'estimation de l'état est fait aléatoirement avec une chance sur deux de se tromper, comme en témoignent les résultats.

Fig. 5.8 Distribution des classes bi-protéiques pour l'évaluation de la classification (lignes de niveau à gauche, visualisation tri-dimensionnelle à droite)



Présentation : données simulées SRM

Ici, nous comparons l'Inversion-Classification (IC) avec les méthodes de classification Naïve Bayes (NB), régression logistique (LR) et *fuzzy c-means* (FCM). On rappelle que les classifieurs IC, NB et LR renvoient les probabilités des classes, FCM un degré d'appartenance aux classes. Ces méthodes concurrentes sont appliquées à des estimations intermédiaires, issues de l'Inversion-Quantification, présentée dans le Ch. 4. La classe estimée est celle avec la plus grande probabilité (IC, NB, LR) correspondant au classifieur MAP, voire celle avec le plus grand degré d'appartenance (FCM), ce qui revient à proposer la fonction de coût 0–1.

Comme précédemment, les méthodes sont comparées selon le risque moyen $R_{c,\mathbf{Y}}(\psi^c)$. Nous l'approchons par l'utilisation de $R = 3000$ échantillons biologiques simulés. Pour démontrer la robustesse de notre approche par rapport au bruit de mesure, nous utilisons cinq niveaux de bruit moyen différents, ainsi que des gains et paramètres chromatographiques variants afin d'imiter des données réelles. Ainsi, nous obtenons des rapports signal à bruit moyen d'environ 0.8, 1.5, 3, 6 et 14 dB.

Nous simulons $\mathsf{M} = 3$ classes, les espèces moléculaires étant données par le schéma $\boxed{P = 2} \xrightarrow{1 \quad 2} \boxed{I = 4} \xrightarrow{1 \quad 3} \boxed{L = 12}$. Les paramètres des classes ont été choisis afin d'obtenir des classes superposées, voir Fig. 5.8.

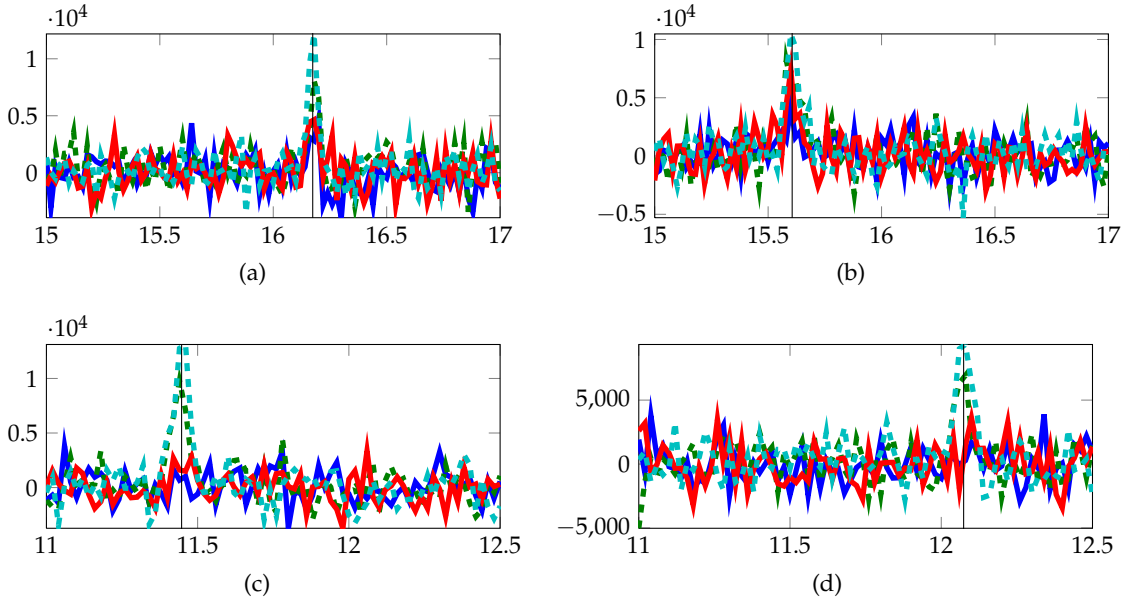
Les résultats de la comparaison sont donnés dans la Tab. 5.3 où nous chiffrons le risque de chaque classifieur associé au coût 0–1 en fonction du rapport signal à bruit (RSB) moyen. La performance croît avec un niveau de bruit décroissant. Comme dans la présentation de résultat précédente, ceci est attendu, quelle que soit la méthode de classification.

Ensuite, on constate que l'IC montre de meilleures performances que les méthodes concurrentes travaillant sur des estimations intermédiaires, quel que soit le rapport signal à bruit. Pour un bruit élevé (RSB de 0.84 dB), 33 % des échantillons test sont mal classifiés par l'IC contre 60 % pour NB, 49 % pour LR et 56 % pour FCM. Pour un RSB moyen de 1.43 dB, le risque moyen de l'Inversion-Classification chute à 0.13, le tiers par rapport aux concurrents. Finalement, nous atteignons une classification presque parfaite en tenant compte des superpositions pour toutes les méthodes de classification en présence d'un très faible bruit.

Discussion : données simulées SRM

Analysons le résultat pour un RSB moyen de 0.84 dB. Le signal est très fortement perturbé, l'information peptidique principale étant noyée dans le bruit, comme démontrent

Fig. 5.9 Donnée simulée fortement perturbée (RSB moyen de 0.51 dB) de quatre peptides, deux transitions natives et deux marquées par peptide, les positions des pics chromatographiques se trouvant à 16.17, 15.61, 11.45 et 12.07 (unité arbitraire) respectivement.



les illustrations dans la Fig. 5.9. Malgré cela, nous obtenons de meilleurs performances par rapport aux autres classifieurs proposés et bien mieux qu'un classifieur aléatoire qui a un risque moyen de $(M-1)/M = 2/3$.

L'écart entre l'IC et les méthodes concurrentes est grand pour un niveau de bruit élevé. Ceci montre que l'IC, grâce à l'information injectée au niveau des paramètres des classes, est performante même proche des limites de détection et de quantification de protéines. Cet écart disparaît naturellement avec un niveau de bruit faible, tous les classifieurs atteignant un risque de 0.05, *i.e.* 5 % de mauvaise estimation de classe. Ce résultat est attendu, surtout suite à l'utilisation de l'Inversion-Quantification pour fournir des estimations protéiques intermédiaires. Que nous n'atteignions pas le classifieur parfait est également attendu. Suite au chevauchement des distributions des classes, la structure du classifieur détermine des coniques de séparation qui coupent les queues des distributions. Ainsi, nous perdons de l'information sur les classes. Cependant, l'analyse des résultats des échantillons dans les queues de classes montre que la probabilité *a posteriori* de la classe estimée est souvent faible et presque équivalente à celle de la classe voisine. Dans une situation de diagnostic, elle serait suffisamment faible pour ne pas se prononcer fermement pour une classe ou une autre, entraînant un traitement d'une maladie inexistante (ou *vice versa*). Il conviendrait de demander une acquisition avec un nouvel échantillon biologique du même individu.

Présentation : données cliniques

Nous arrivons maintenant au sommet de cette étude : la confrontation de la méthode d'Inversion-Classification à des données réelles cliniques. Les données sont issues du jeu du cancer colorectal (voir Sect. 3.6) et sont toutes étiquetées par leur classe d'appartenance initiale, *i.e.* cas (dans les figures illustré en rouge) ou contrôle (illustré en bleu). Suite à une étude préalable, six protéines potentiellement biomarqueurs ont été avancées dont seulement deux sont suffisamment différentiantes selon nos analyses : la LFABP et

Tab. 5.3 Évolution du risque des classifieurs en fonction du niveau moyen de bruit de mesure

RSB	IC	NB	LR	FCM
0.84	0.33	0.60	0.49	0.56
1.47	0.13	0.42	0.41	0.39
3.04	0.08	0.19	0.12	0.15
6.11	0.04	0.06	0.05	0.04
14.30	0.04	0.05	0.05	0.04

la Protéine X.

Nous mettons en place une approche de validation croisée aléatoire pour l'analyse sur les données cliniques (voir Sect. 3.6.4 pour la présentation de la cohorte). Soit $\mathcal{P}$ l'ensemble de tous les individus, quelle que soit leur classe, et $\mathcal{P}_+$ et $\mathcal{P}_-$ les sous-cohortes cas et contrôle respectivement. Nous divisons les sous-cohortes en sélectionnant aléatoirement le même nombre d'individus que nous utilisons pour la validation, et soit V ce nombre par sous-cohorte. Nous notons les sous-cohortes de validation $\mathcal{P}_+^V$ et $\mathcal{P}_-^V$ selon la provenance cas ou cohorte. Les sous-cohortes d'apprentissage contiennent alors $\text{card}(\mathcal{P}_+^V) - V$ et $\text{card}(\mathcal{P}_-^V) - V$ individus à partir desquels nous estimons les paramètres de classe respectivement.

Une telle approche nous permet de nous affranchir d'un sur-apprentissage que l'on obtient en classant les mêmes individus sur lesquels on a entraîné le classifieur. Nous mettons cette procédure en boucle de $R = 20$ répétitions, en choisissant un nombre d'individus de validation $V = 10$, obtenant ainsi des sous-cohortes d'apprentissage contenant environ 90 % des individus par classe, soit 88 pour la sous-cohorte « cas » et 100 pour la sous-cohorte « contrôle ».

La Tab. 5.4 montre le tableau de fréquences empiriques relatives sur $R = 20$ essais des quatre méthodes de classification mises en concurrence. Le nombre de vrais négatifs (*i.e.* vérité = cas, estimation = cas) est élevé pour les quatre méthodes, le FCM donnant de loin les moins bons résultats (60 % de vrais négatifs), le classifieur IC montrant les meilleurs résultats sans erreur (100 %). Ainsi, cette approche affiche un taux de faux positif nul : aucun individu « cas » n'a été classé dans la classe « contrôle ».

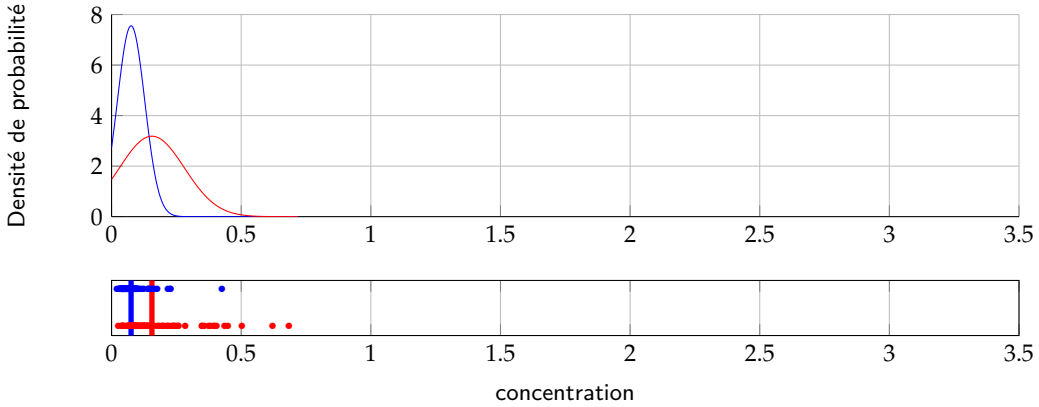
Le nombre de vrais positifs est cependant moins élevé : le FCM est ici le plus performant (80 %), NB le moins performant (55 %). Le risque moyen des méthodes que l'on peut également déduire de ce tableau se chiffre respectivement pour chaque classifieur à $\mathrm{R}_{C,\mathbf{Y}}^{\mathrm{IC}} = 0.095$, $\mathrm{R}_{C,\mathbf{Y}}^{\mathrm{NB}} = 0.1125$, $\mathrm{R}_{C,\mathbf{Y}}^{\mathrm{LR}} = 0.10$ et $\mathrm{R}_{C,\mathbf{Y}}^{\mathrm{FCM}} = 0.15$. L'IC obtient le risque moyen minimum.

Discussion : données cliniques

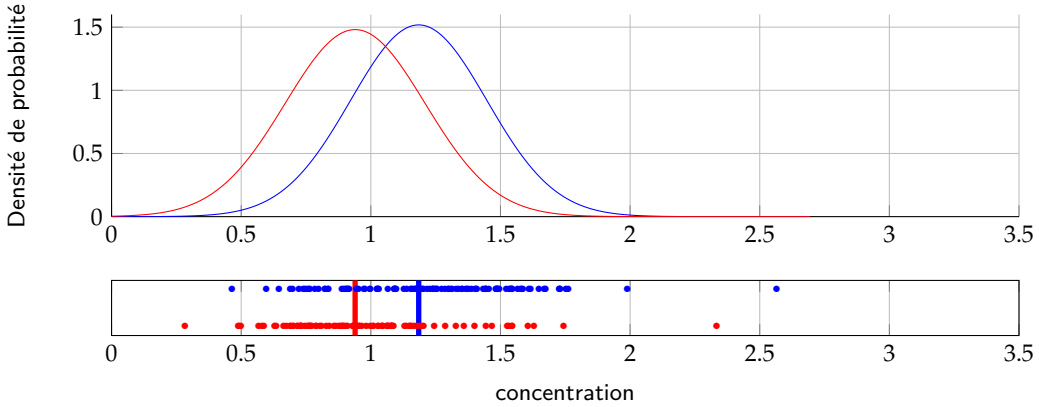
Les méthodes ont des performances similaires (à l'exception du FCM). Une des raisons pour cela est sans doute le fait que l'estimation intermédiaire est fournie par l'Inversion-Quantification. Celle-ci intègre déjà de la variabilité technologique dans son estimation qui est robuste par rapport au bruit de mesure, comme nous l'avons démontré précédemment. Les résultats auraient été certainement encore plus favorables pour l'IC si l'estimation intermédiaire des méthodes concurrentes aurait été fournie par une méthode de quantification de l'état de l'art moins robuste au bruit.

Ensuite, les quatre méthodes affichent un taux de faux négatif relativement élevé, résultant en des risques $\mathrm{R}_{\mathbf{Y}|C=\text{cas}}$ entre 0.2 et 0.45. En considérant les distributions, marginales dans la Fig. 5.10 et jointe dans la Fig. 5.11, on voit que les classes sont sensiblement superposées. Le choix des biomarqueurs est donc très important ! Pour cette comparai-

Fig. 5.10 Densités de probabilité marginales de la concentration des protéines considérées dans le jeu de données cliniques pour les sous-cohortes contrôle (bleu) et cas (rouge) entières en haut de chaque sous-figure ; en dessous, la distribution de points marginale représentant les estimations de concentration (points •) pour chaque classe, le trait vertical correspond à la moyenne des classes (mêmes couleurs). L'unité de la concentration est relative à la concentration de l'échantillon CQ3.



(a) Protéine 1

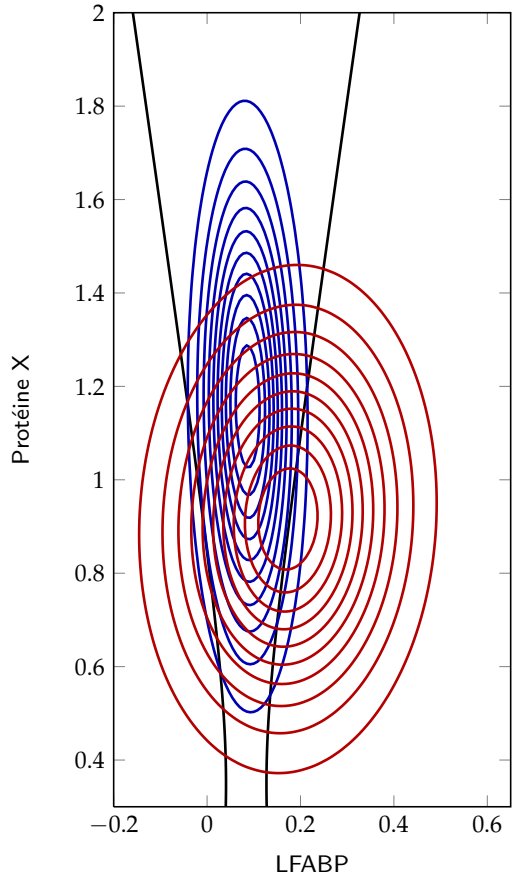


(b) Protéine 2

son, nous avons déjà choisi dans une liste de six candidats les deux protéines les plus différentiantes, la protéine LFABP montrant un effet de sur-expression, la Protéine X une sous-expression de la concentration protéique quand la maladie est présente. Ce choix a été effectué manuellement, une sélection automatique étant l'une des perspectives du travail.

Considérons maintenant les performances du classifieur Inversion-Classification. Son taux de faux négatifs (38 %), *i.e.* classement d'un individu malade dans la classe contrôle, est relativement élevé. Si nous comparons avec les autres classifieurs probabilistes (NB, LR), nous remarquons une tendance similaire. Cependant, nous affichons un taux de vrais négatifs de 100 %, tous les individus sains ont été correctement classés dans la classe contrôle. Autrement dit, le taux de faux positifs est nul ce qui est la valeur la plus faible parmi les méthodes étudiées. Les classifieurs probabilistes affichent de bons résultats quant aux vrais négatifs, contrairement au FCM. La réponse est peut-être simplement donnée par le fait que le FCM semble avoir une « préférence » pour une estimation en « cas », alors que les autre méthodes « préfèrent » l'estimation en « contrôle ». En effet, quand on considère les résultats du FCM, on s'aperçoit qu'il a à la fois le plus grand

Fig. 5.11 Densité jointe des protéines $1 \div 2$ du jeu de données cliniques (lignes de niveau) ; bleu : contrôle, rouge : cas. En noir, l'hyperbole de séparation des classes déterminée par l'IC.



taux de faux positifs (40 %) et le plus grand taux de vrais positifs (80 %), inversement aux autres classifieurs.

Quelle est la raison pour cette préférence ? Pour comprendre la raison au moins pour les classifieurs probabilistes, considérons la distribution marginale *a posteriori* des classes, Fig. 5.11. Les paramètres appris sont $\boldsymbol{m}_+ = [0.173, 0.916]^{\mathrm{T}}$, $\boldsymbol{\Gamma}_+ = \left(\begin{smallmatrix} 47.475 & -1.7408 \\ -1.7408 & 16.282 \end{smallmatrix}\right)$ et $\boldsymbol{m}_- = [0.0869, 1.157]^{\mathrm{T}}$, $\boldsymbol{\Gamma}_- = \left(\begin{smallmatrix} 293.08 & 2.88 \\ 2.88 & 11.23 \end{smallmatrix}\right)$ respectivement pour les classes « cas » et « contrôle ». En suivant [85], on peut déterminer que la conique de séparation correspond à une hyperbole : le déterminant de la différence des précisions est négatif, le logarithme du produit des valeurs propres du rapport des précisions est négatif et, de plus, supérieur au terme additif constant issu du rapport des deux lois normales considérées, et enfin la matrice de rotation associée correspond à une rotation de 179 ° (p. 17 de ladite référence). Nous avons ajouté le tracé de l'hyperbole en noir dans la Fig. 5.11. On déduit de cette analyse que si une concentration protéique se trouve géométriquement « entre les branches » de l'hyperbole, la classe estimée est « contrôle » ; sinon, l'estimation est « cas ». Cela se traduit par le fait que nous avons observé : nous obtenons un taux de faux positifs nul car toutes les concentrations protéiques de la sous-cohorte « contrôle » se trouve entre les branches de l'hyperbole, mais nous avons un taux de faux négatif non nul car l'intersection de l'intérieur des branches de l'hyperbole avec la distribution de la concentration protéique de la sous-cohorte « cas » est non vide. Les autres méthodes probabilistes affichent un résultat similaire : le classifieur NB travaille avec la même conique, mais avec des distributions normales séparables ; le classifieur LR impose quant à

Tab. 5.4 Tableaux de fréquences empiriques relatives de la classification sur les données cliniques en mettant en place une validation croisée aléatoire

(a) IC. Risque moyen associé : $R^{IC}_{C,\mathbf{Y}} = 0.095$

vérité \ estimation	contrôle	cas
contrôle	1	0
cas	0.38	0.62

(b) NB. Risque moyen associé : $R^{NB}_{C,\mathbf{Y}} = 0.1125$

vérité \ estimation	contrôle	cas
contrôle	0.985	0.015
cas	0.45	0.55

(c) LR. Risque moyen associé : $R^{LR}_{C,\mathbf{Y}} = 0.10$

vérité \ estimation	contrôle	cas
contrôle	0.95	0.05
cas	0.35	0.65

(d) FCM. Risque moyen associé : $R^{FCM}_{C,\mathbf{Y}} = 0.15$

vérité \ estimation	contrôle	cas
contrôle	0.6	0.4
cas	0.2	0.8

lui l'homoscédasticité des lois conduisant à une droite de séparation.

Finalement, l'Inversion-Classification affiche le risque moyen le moins élevé (sur vingt répétitions ...) et semble moins « tiré » vers l'une ou l'autre possibilité d'estimation. En effet, l'estimateur bayésien est celui qui minimise en théorie le coût *a posteriori* et, par conséquent en moyennant sur les données, le risque moyen. Ceci est confirmé par cette analyse pour cette campagne.

5.5 Conclusion

À travers ce chapitre, nous avons introduit un moyen de diagnostic ou d'aide au diagnostic. Pour cela, nous avons défini le terme « classification » et avons proposé de distinguer deux tâches : l'« apprentissage » et l'« estimation de classe ». Ensuite, nous avons adapté la structure hiérarchique pour ces tâches en ajoutant un niveau classe/paramètres de classe en début de chaîne. Nous avons justifié l'utilisation des méthodes statistiques bayésiennes qui sont en l'occurrence bien adaptées à de telles structures.

Après avoir exposé le développement des méthodes pour réaliser ces tâches et les mises en œuvres associées, nous avons discuté les résultats respectifs. Nous avons vu que l'apprentissage affichait une robustesse par rapport aux variabilités et pouvait être réalisé sur des cohortes de taille limitée. Quant à la classification, nous avons vérifié que notre approche donnait des taux de mauvaise classification faibles. Ceci correspond notamment au risque moyen le moins élevé.

Ensuite, nous avons analysé les résultats de la classification sur des données réelles et ont conclu qu'ils étaient satisfaisant : nous avons atteint de bons apprentissages et des taux de mauvaise classification faibles. Nous avons pu travailler à la fois sur des données simulées Full-MS et SRM, permettant de confronter les estimations à des valeurs vraies, et des données réelles SRM, permettant cette fois de confronter les méthodes d'inversion à des observations qui ne suivent pas strictement le modèle physique.

Alg. 5.1 Résumé de l'algorithme d'estimation des paramètres d'une classe

Entrées: $\mathbf{Y}_{1:N}$, paramètres *a priori* for $\boldsymbol{\pi}$, $\boldsymbol{x}$, $\boldsymbol{\kappa}$, $\boldsymbol{\theta}^{\text{inst}}$
Nombre maximal d'itérations : K^{M}.
Nombre minimal d'itérations de temps de chauffe : K_0.
Nombre d'échantillons pris en compte : K.

fonction $\hat{\boldsymbol{\pi}} = \textsc{Apprentissage}(\mathbf{Y}_{1:N})$
  Initialisation: $k = 0$, initialiser les paramètres

  tant que $k \leq K^{\text{M}}$ **faire**
    $k \leftarrow k + 1$
    pour $n = 1, \ldots, N$ **faire**
      échantillonner $(\boldsymbol{\theta}_n^{\text{inst}})^{(k)} \sim \mathrm{p}(\boldsymbol{\theta}^{\text{inst}} \mid \mathbf{Y}_n, \boldsymbol{\kappa}_n^{(k-1)})$
        ▷ Produit non standard de distributions. Étape de Gibbs hybride.
      échantillonner $\boldsymbol{\kappa}_n^{(k)} \sim \mathrm{p}(\boldsymbol{\kappa} \mid \mathbf{Y}_n, (\boldsymbol{\theta}_n^{\text{inst}})^{(k)}, \boldsymbol{x}_n^{(k-1)}$
        ▷ Distribution normale multivarée. Échantillonnage explicite.
      échantillonner $\boldsymbol{x}_n^{(k)} \sim \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\kappa}_n^{(k)}, \boldsymbol{\pi}^{(k-1)})$
        ▷ Distribution normale multivarée. Échantillonnage explicite.
    fin pour

    échantillonner $\boldsymbol{\pi}^{(k)} \sim \mathrm{p}(\boldsymbol{\pi} \mid \boldsymbol{x}_{1:N}^{(k)})$
      ▷ Distribution normale-wishartienne. Échantillonnage explicite.
    si Critère d'arrêt du temps de chauffe satisfait **et** $k \geq K_0$ **alors**
      Poser $K^{\text{M}} \leftarrow k + K$ et $\mathcal{K} := \{k+1, \ldots, k+K\}$
    fin si
  fin tant que

  retourner $\hat{\boldsymbol{\pi}} = \dfrac{1}{K} \sum_{k \in \mathcal{K}} \boldsymbol{\pi}^{(k)}$
fin fonction

Alg. 5.2 Résumé de l'algorithme d'estimation de la classe

Entrées: $\mathbf{Y}$, distribution catégorielle pour C, paramètres $\boldsymbol{\pi}$ pour la distribution de $\boldsymbol{x}$, paramètres *a priori* pour $\boldsymbol{\kappa}$, $\boldsymbol{\theta}^{\text{inst}}$
Nombre maximal d'itérations : K^{M}.
Nombre minimal d'itérations de temps de chauffe : K_0.
Nombre d'échantillons pris en compte : K.

1 **fonction** $\hat{\mathrm{c}} = \textsc{Classement}(\mathbf{Y})$
2 **Initialisation:** $k = 0$, initialiser les paramètres

3 **tant que** $k \leq K^{\text{M}}$ **faire**
4 $k \leftarrow k + 1$
5 échantillonner $(\boldsymbol{\theta}^{\text{inst}})^{(k)} \sim \mathrm{p}(\boldsymbol{\theta}^{\text{inst}} \mid \mathbf{Y}, \boldsymbol{\kappa}^{(k-1)})$
▷ Produit non standard de distributions. Étape de Gibbs hybride.
6 échantillonner $\boldsymbol{\kappa}^{(k)} \sim \mathrm{p}(\boldsymbol{\kappa} \mid \mathbf{Y}, (\boldsymbol{\theta}^{\text{inst}})^{(k)}, \boldsymbol{x}^{(k-1)})$
▷ Distribution normale multivarée. Échantillonnage explicite.
7 échantillonner $\boldsymbol{x}^{(k)} \sim \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\kappa}^{(k)}, \boldsymbol{\pi}^{(k-1)})$
▷ Distribution normale multivarée. Échantillonnage explicite.

8 échantillonner $\mathrm{c}^{(k)} \sim \Pr(\mathrm{c} \mid \boldsymbol{x}^{(k)})$
▷ Distribution catégorielle. Échantillonnage explicite.
9 **si** Critère d'arrêt du temps de chauffe satisfait **et** $k \geq K_0$ **alors**
10 Poser $K^{\text{M}} \leftarrow k + K$ et $\mathcal{K} := \{k+1, \ldots, k+K\}$
11 **fin si**
12 **fin tant que**

13 **pour** $\mathsf{m} = 1, \ldots, \mathsf{M}$ **faire**
14 $\mathcal{K}^{\mathsf{m}} = \left\{ k \in \mathcal{K} \mid \mathrm{c}^{(k)} = \mathrm{c}^{\mathsf{m}} \right\}$
15 $\Pr(C = \mathrm{c}^{\mathsf{m}} \mid \mathbf{Y}) = \frac{\operatorname{card} \mathcal{K}^{\mathsf{m}}}{K}$
16 **fin pour**
17 **retourner** $\hat{\mathrm{c}} = \underset{\mathsf{m}=1,\ldots,\mathsf{M}}{\operatorname{argmax}} \Pr(C = \mathrm{c}^{\mathsf{m}} \mid \mathbf{Y})$
18 **fin fonction**

6 Conclusions et perspectives

6.1 Conclusions

Dans ce document, nous avons étudié la méthodologie des problèmes inverses couplée au cadre des statistiques bayésiennes en protéomique. Nous avons mis en avant notamment le problème de la **quantification** de protéines et du diagnostic à travers la **classification** d'un échantillon biologique incluant une phase préalable d'**apprentissage** sur une cohorte d'échantillons.

Nous avons développé ces applications pour une chaîne d'analyse comportant un chromatographe liquide et un spectromètre de masse soit en mode « Full-MS », soit en mode « *Selected Reaction Monitoring* », avec calibrage interne et externe. Nous avons proposé une modélisation directe de l'acquisition de données, en se basant sur les travaux de Strubel [1] pour le mode Full-MS et sur les travaux de Gerfault et al. [117] pour le mode SRM, et avons constaté qu'elle reposait sur un modèle direct hiérarchique. Cette hiérarchie se propage ensuite dans la modélisation probabiliste et, *in fine*, dans l'inversion.

Ce travail a montré l'apport de l'inversion bayésienne d'un modèle hiérarchique en protéomique. Par rapport à la quantification, cette méthodologie permet de travailler sur le paramètre qui est vraiment d'intérêt en protéomique : la protéine. Au Ch. 4, nous avons montré sur l'Inversion-Quantification que l'on arrive à des estimations de concentrations protéiques avec une dispersion faible à la fois pour des données simulées et pour des données réelles, malgré un rapport signal sur bruit parfois défavorable. Sur les données du cancer colorectal, nous avons montré une bonne corrélation avec les tests ELISA disponibles, ainsi que le caractère « auto-sélectif » de la méthode quand on est en présence d'une mesure trop perturbée. Ces analyses de résultats nous ont permis de valider l'approche Inversion-Quantification.

Ensuite, au Ch. 5 sur l'Inversion-Classification, nous avons étendu le problème d'estimation de paramètres continus au problème d'estimation de paramètres discrets. Pour la classification, nous avons séparé le problème en deux sous-problèmes : apprentissage et estimation de la classe. Nous avons évalué le premier avec des données simulées, montrant qu'on arrive à approcher la distribution de la population avec celle issue de l'apprentissage, *modulo* des incertitudes sur les estimations. Ensuite, nous avons réalisé l'évaluation de l'estimation de la classe sur des cohortes simulées et sur une cohorte réelle. Dans tous les cas, nous avons souligné le bon fonctionnement de la méthode, et avons pu montrer qu'elle améliorait les résultats par rapport à l'état de l'art travaillant sur des estimations intermédiaires, menant à un nombre moins élevé de fausses détections à la fois dans le sens « faux positifs » que « faux négatifs », validant ainsi l'Inversion-Classification.

Pour la mise en œuvre algorithmique, nous nous sommes basés sur l'utilisation de l'échantillonnage stochastique de la loi *a posteriori* jointe reposant sur la méthode *Monte Carlo Chaîne de Markov*. À l'intérieur de cette méthode, nous avons adopté une structure de Gibbs avec un algorithme de Gibbs hybride, combinant au moins une étape de Metropolis-Hastings et des générateurs d'échantillons explicites, pour calculer les estimateurs et les incertitudes associées.

Par rapport à l'état de l'art, nos développements ont l'avantage de nécessiter peu ou pas d'ajustements manuels car tous les paramètres sont estimés ce qui donne à notre méthode son caractère auto-adaptatif ; grâce à l'utilisation de calibrants internes, elle est également auto-calibrante. Nous avons également montré en quantification et en classification que l'utilisation du cadre bayésien se justifie d'autant plus que le rapport signal sur bruit se dégrade.

Le travail présenté ici prolonge les travaux menés dans l'équipe PROTIS depuis plusieurs années, notamment à travers la thèse de Strubel [1]. Nous avons atteint quelques perspectives proposées par l'auteur :

1. *« Enfin, nous pouvons améliorer les informations a priori injectées. Nous pourrons mieux modéliser la gamme de concentration attendue [. . .] »*
   L'apprentissage avait été introduit avec le but de connaître la distribution de la concentration des biomarqueurs dans une classe donnée. Cette connaissance peut servir d'*a priori* pour la quantification et restreint ainsi la gamme de valeurs possibles. Nous avons proposé l'utilisation comme *a priori* pour la classification, renseignant sur les classes. C'est un des éléments clé de la méthode que nous avons présentée.
2. *« De même, nous pouvons envisager de traiter en une fois plusieurs expériences. »*
   Nous avons montré une méthode d'apprentissage qui traite plusieurs données à la fois. Nous ne sommes pas obligés de retourner uniquement les paramètres de classe que nous avons calculés, mais nous pouvons – comme nous l'avons remarqué – calculer les estimateurs pour chacune des expériences ou, s'il s'agit d'un N-tuple réplicat du même échantillon de base, donner la concentration moyenne avec son incertitude.
3. *« De plus, nous pouvons noter que la quantification n'est habituellement pas le but final du traitement. En effet, ces concentrations seront utilisées pour découvrir de nouveaux marqueurs ou pour réaliser un diagnostic. »*
   Nous avons proposé un moyen d'aide au diagnostic reposant la classification d'un nouvel échantillon. La découverte de nouveaux marqueurs s'inscrit également dans nos perspectives.
4. *« Nous pourrons [. . .] essayer un échantillonneur de Metropolis-Hastings à marche aléatoire. »*
   C'est en effet cet échantillonneur que nous avons mis en œuvre avec succès et satisfaction. Il semble être un bon compromis entre les différentes approches de Metropolis-Hastings.
5. *« Nous n'avons pour l'instant utilisé qu'un peptide pour effectuer la quantification, or la méthode PSAQ et notre méthode de traitement permettent d'utiliser tous les peptides de la protéines. »*
   Les pics peptidiques sur des données LC-Full-MS réelles peuvent être très espacés ce qui demande de mettre en mémoire une très grande matrice ; cependant, nous avons utilisé avec succès plusieurs peptides sur des données simulées. Le mode SRM s'affranchit des grandes zones d'« intérêt nul » que l'on peut observer en Full-MS en ciblant les masses des molécules d'intérêt. L'estimation à partir de plusieurs peptides est donc moins « gourmande » en besoin mémoire et en temps de calcul. Toutes les données traitées (simulées ou réelles) comportent plusieurs peptides par protéine, et tous ont été utilisés pour l'estimation des concentrations protéiques (modulo les calibrages interne et externe). [↓1]

1. Notons cependant que G. Strubel a étendu sa méthode « mono-peptidique » à une méthode « multi-peptidique » pour la soumission de la révision de [118] en utilisant des imagettes de l'acquisition globale.

6. *« Peut-on utiliser la méthode sur d'autres spectromètres de masse et d'autres méthodes de marquage isotopique ? »*
   La réponse est oui, comme nous l'avons démontré dans ce document par l'utilisation du mode SRM avec le marquage AQUA. Les modifications nécessaires pour arriver à notre fin contiennent l'adoption d'une structure hiérarchique, la modélisation de la chaîne d'analyse suivant cette hiérarchie et l'adaptation des méthodes d'inversion à l'utilisation de plusieurs traces de données (au lieu d'une grande matrice de données).

6.2 Perspectives

Quand une thèse démarre, trois ans paraissent longs, et le doctorant peut se dire qu'il aura tout son temps pour investiguer des pistes pour faire et parfaire son travail. Hélas (ou fort heureusement ? !), ce n'est souvent pas vrai car notamment dans les derniers mois les idées d'amélioration, de précision, d'approfondissement, ... surgissent. C'est également le cas ici : nous avons proposé un cadre bayésien/problèmes inverses pour la quantification et la classification en protéomique, et à la fin de ce travail nous pouvons proposer une liste (non exhaustive) de thèmes de recherche pour poursuivre nos travaux.

Modèle

Par rapport aux modèles utilisés pour les données et leur construction, nous chercherons les réponses aux problèmes suivants liés à la quantification. Le modèle chromatographique utilisé est approximatif. En réalité, les pics chromatographiques observés ne sont pas gaussiens, mais ont une légère traînée ou asymétrie [67]. Les modèles bigaussiens [93] ou des gaussiens modifiés exponentiellement [94, 95] sont une meilleure approximation, les modèles à base de *splines* sont également utilisés dans la littérature [96], et enfin les modèles issus des équations différentielles approchent encore plus la réalité. Qu'apportent ces changements, ces précisions du modèle par rapport au modèle gaussien « imparfait » ? Quels sont les inconvénients ?

Le modèle chromatographique présenté est linéaire. Nous avons même fixé la largeur des pics à une valeur calibrée en amont de l'expérience pour l'inversion des données Full-MS. Or, les expériences montrent que la largeur des pics dépend du temps de rétention et de la concentration de la molécule.

Dans les données réelles, le bruit n'est pas gaussien de moyenne nulle comme nous le supposons dans la modélisation. Même si on peut supposer l'erreur de modélisation distribuée ainsi, l'erreur de mesure des instruments considérés étant non négative a une autre distribution. Qu'est-ce qu'on gagne en terme d'estimation en ayant un modèle plus « réaliste » du bruit, comparé à ce que l'on perd (conjugaison des lois, échantillonnages explicites, augmentation de la demande et du temps de calcul...) ?

Même si l'analyse de données simulées dans [118] a montré que les objections précédentes sont négligeables, leur étude peut être bénéfique pour la compréhension des phénomènes ou pour l'investigation d'autres approximations.

L'analyse des données pour la quantification a également montré que certaines traces sont très ou trop perturbées pour permettre leur utilisation dans l'estimation. Nous avons certes amorti l'influence de ces traces en « pondérant » par la qualité des sources, mais elles dégradent toujours la performance de la quantification. Quel paramètre peut nous renseigner sur une possible élimination de la trace ?

C'est d'ailleurs ce qu'il propose dans ses perspectives de thèse.

Apprentissage, robustesse, et étiquettes manquantes

Les étiquettes des données utilisées pour la classification n'ont pas été mises en cause. Pourtant, comme nous l'avons dit, il n'y a pas d'« individu sain » ; il s'agit uniquement d'un individu qui ne montre pas les symptômes visibles depuis l'extérieur de la maladie considérée. Son profil protéique cependant peut déjà être modifié. Ainsi, l'étiquette attribuée peut être fausse. Avec l'utilisation des méthodes bayésiennes, on peut ajouter un indicateur de fiabilité sur l'étiquette, puis réattribuer une nouvelle étiquette lors de l'apprentissage. Ceci nécessite de faire un apprentissage global (pas classe par classe comme nous le faisons actuellement car la séparation ne peut plus être satisfaite). On peut s'inspirer de l'article [88] ou plus généralement des approches « apprentissage semi-supervisé », « partitionnement », « données manquantes ».

Algorithme

Au niveau du développement algorithmique, nous travaillons actuellement avec MATLAB pour simuler une chaîne MCMC qui propose des échantillons de la loi cible. La tendance de la communauté bayésienne semble cependant aller vers l'utilisation des méthodes bayésiennes variationnelles qui sont réputées plus rapides que les méthodes d'échantillonnage stochastique. Elles fournissent une approximation séparable de la loi *a posteriori* donnant accès une expression analytique, les estimateurs étant pris ensuite sur cette approximation. Une étude sur son application à la protéomique pourrait être menée pour quantifier l'apport par rapport à l'existant et pour permettre de comparer les résultats, temps de calcul, investissements de développement, *et cetera*.

La validation faite dans ce rapport repose sur des études de petite dimension (petit nombre de protéines, cohortes d'effectifs relativement faibles, cohortes bi- ou triclasses). Les contraintes de la puissance des tests poussent à travailler en grandes dimensions. L'augmentation du nombre de protéines, d'effectifs ou de classes demande des adaptations algorithmiques afin d'être moins gourmand en termes de temps de calcul et de mémoire.

Évaluation

L'évaluation de la quantification des données cliniques a été faite en comparant nos résultats avec une autre estimation fournie par les tests ELISA. Cependant, ne connaissant pas l'incertitude et les erreurs de ces derniers, l'analyse de la corrélation n'est pas suffisamment concluante et peut mener à des déductions erronées. C'est la raison pour laquelle notre partenaire bioMérieux de procède sur le projet BHI-PRO à une nouvelle campagne d'acquisition de mesures, comportant également une répétition des tests ELISA. Ainsi, nous pourrons estimer leur variabilité, calculer les coefficients de corrélation entre les méthodes de mesure par spectrométrie de masse et les méthode de mesure ELISA [72] et consolider nos conclusions sur la comparaison des différentes méthodes. Sur BHI-PRO nos collègues du LBS développent une méthodologie d'analyse statistique reposant sur des techniques de régression associées à des variables incertaines et des techniques d'étude de corrélations ou d'analyse de varaince. L'évaluation nécessite aussi de mettre en place un plan d'expérience adapté.

Il serait également intéressant de poursuivre la comparaison des méthodes que nous avons présentées ici avec les méthodes d'autres équipes. Ainsi, nous pourrons par exemple confronter l'Inversion-Quantification au logiciel libre Skyline [112].

Sélection de biomarqueurs

Le projet ANR BHI-PRO s'inscrit parfaitement dans ce que nous venons de conclure : les questions classification (incluant l'apprentissage) et quantification sont au cœur de ce projet. Le projet s'intéresse à la question de la découverte et la validation de biomarqueurs. Nous travaillons en collaboration avec les équipes de biostatistiques du LBS (Laboratoire Biostatistique Santé, Lyon) et du CLIPP (CLInical Proteomic Platform, Dijon). À partir d'un protéome déjà identifié comme candidat différentiant, on veut répondre aux questions « Est-ce que ce protéome est vraiment différentiant, est-ce qu'un sous-protéome est différentiant, est-ce qu'on peut en réduire la taille ? » Des méthodes pour répondre à cette question sont vivement débattues dans le projet BHI-PRO. On peut par exemple partir de la question de la séparation des classes par l'ensemble des protéines candidates (« Est-ce que l'histogramme multidimensionnel possède un ou deux modes ? »), et affiner ensuite de manière à trouver un sous-ensemble différentiant de taille réduite. Une autre piste possible est la recherche d'un ensemble efficace et le plus parcimonieux possible par des méthodes de choix de modèle impliquant une pénalisation de la complexité. L'article de revue [97] expose des méthodes de sélection de variables/modèle dans une stratégie bayésienne et leurs implantations algorithmiques et les rapports de stage [85, 98] proposent des éléments de réponse à cette question.

Spectrométrie de masse en mode MRM^3

Dans cette thèse, un mode de spectrométrie étudié repose sur la fragmentation d'un peptide. La double-fragmentation d'une molécule que l'on observe dans le mode MRM^3 d'un spectromètre de masse apporte de la spécificité supplémentaire. L'adaptation des modèles ou l'utilisation des algorithmes existants permettant l'exploitation des données de ce mode de spectrométrie fait partie des objectifs du projet BHI-PRO et des perspectives de ce travail.

A Calculs détaillés

A.1 Classifieur utilisant une fonction de coût pondérée

Soit $c^\star \in \mathcal{C} = \{c^1, c^2\}$ la classe de l'échantillon $\mathbf{Y}$, et soit $\psi^c : \Omega_\mathbf{Y} \to \mathcal{C}$ le classifieur qui associe à $\mathbf{Y}$ une estimation de classe $\psi(\mathbf{Y}) = \hat{c}$. Nous avons vu dans Sect. 2.6.3 que le classifieur associé à la fonction de coût 0–1 est l'estimateur Maximum A Posteriori. Cette fonction associe le coût 0 à une bonne estimation de classe et le coût 1 à une mauvaise estimation quelle que soit la classe.

Dans ce qui suit, nous voulons déduire la règle du classifieur associé à une fonction de coût pondéré, *i.e.* une fonction qui n'associe pas le même coût à une mauvaise estimation dans c^1 que dans c^2. Pour cela, nous réécrivons le critère du MAP et l'étendons dans la suite.

En effet, le critère de classification MAP est donné par $\psi^c(\mathbf{Y}) = \operatorname{argmax}_{c\in\mathcal{C}} \Pr(C = c \mid \mathbf{Y})$. Soit, sans perte de généralité, $\psi^c(\mathbf{Y}) = c^1$. La probabilité *a posteriori* de la première classe étant maximale et l'ensemble des classes étant constitué de deux éléments, ceci est équivalent à $\Pr(C = c^1 \mid \mathbf{Y}) > \Pr(C = c^2 \mid \mathbf{Y})$. Ces deux lois possédant le même facteur de normalisation $p(\mathbf{Y})$ en appliquant la règle de Bayes, nous écrivons $p(\mathbf{Y} \mid C = c^1) \Pr(C = c^1) > p(\mathbf{Y} \mid C = c^2) \Pr(C = c^2)$ et obtenons

$$\frac{p(\mathbf{Y} \mid C = c^1)}{p(\mathbf{Y} \mid C = c^2)} > \frac{\Pr(C = c^2)}{\Pr(C = c^1)} = \left(\frac{\Pr(C = c^1)}{\Pr(C = c^2)}\right)^{-1}. \tag{A.1}$$

Autrement dit, si c^1 maximise la loi *a posteriori* de la classe, alors le rapport des vraisemblances est supérieur à l'inverse cote des probabilités *a priori*.

Donc, le classifieur MAP peut s'écrire de manière équivalente à ce que nous avons introduit dans Sect. 2.6.3 sous la forme suivante :

$$\psi(\mathbf{Y}) = \begin{cases} c^1 & \text{si } V_\mathbf{Y}(c) > T \\ c^2 & \text{si } V_\mathbf{Y}(c) \leq T \end{cases} \tag{A.2}$$

où $V_\mathbf{Y}(c) = \frac{p(\mathbf{Y}|C=c^1)}{p(\mathbf{Y}|C=c^2)}$ est le rapport des vraisemblances et $T = \frac{\Pr(C=c^2)}{\Pr(C=c^1)}$ l'inverse cote *a priori*.

Soit maintenant la fonction de coût suivante :

$$L(c, \psi^c(\mathbf{Y})) = \begin{cases} 0 & \text{si } \psi^c(\mathbf{Y}) = c, \\ \ell_1 & \text{si } \psi^c(\mathbf{Y}) \neq c^1 \text{ alors que } C = c^1, \\ \ell_2 & \text{si } \psi^c(\mathbf{Y}) \neq c^2 \text{ alors que } C = c^2. \end{cases} \tag{A.3}$$

En mots, une bonne décision ne coûte évidemment rien. Si la bonne classe est c^1, une mauvaise classification coûte ℓ_1 ; si la bonne classe est c^2, une mauvaise classification coûte ℓ_2.

Le raisonnement permet de proposer un nouveau seuil d'inverse cote *a priori* qui est pondéré par la cote des coûts : $T_\ell = \frac{\Pr(C=c^2)}{\Pr(C=c^1)} \frac{\ell_2}{\ell_1}$ interprétant ainsi les coûts comme des connaissances et pondérant la distribution *a priori*.

Si on remplace T de l'Éqn. (A.2) par ce nouveau seuil T_ℓ, on obtient

$$\begin{aligned}
\frac{\mathrm{p}(\mathbf{Y} \mid C = \mathrm{c}^1)}{\mathrm{p}(\mathbf{Y} \mid C = \mathrm{c}^2)} &> \frac{\Pr(C = \mathrm{c}^2)}{\Pr(C = \mathrm{c}^1)} \frac{\ell_2}{\ell_1} \\
\iff \quad \frac{\Pr(C = \mathrm{c}^1 \mid \mathbf{Y})}{1 - \Pr(C = \mathrm{c}^1 \mid \mathbf{Y})} &> \frac{\ell_2}{\ell_1} \\
\iff \quad \Pr(C = \mathrm{c}^1 \mid \mathbf{Y}) &> \frac{\ell_2}{\ell_1 + \ell_2}.
\end{aligned}$$

Ainsi, pour que la classe c^1 soit choisie, sa probabilité *a posteriori* doit excéder $\ell_2/(\ell_1 + \ell_2)$. Si ce n'est pas le cas, la classe c^2 dont la probabilité *a posteriori* est automatiquement plus grande que $\ell_1/(\ell_1 + \ell_2)$, est choisie. On note aussi que l'utilisation de la fonction de coût 0–1 est un cas spécial avec $\ell_1 = \ell_2$ que l'on peut poser arbitrairement à 1.

A.2 Erreur quadratique

Soit $\mathbf{Y} \sim \mathrm{p}(\mathbf{Y} \mid \theta)$ une variable aléatoire sous $\mathrm{p}(\mathbf{Y} \mid \theta)$, le paramètre θ étant fixé. Soit $\psi(\cdot) : \Omega_\mathbf{Y} \to \Theta$ un estimateur qui associe à une mesure $\mathbf{Y} \in \Omega_\mathbf{Y}$ une valeur $\psi(\mathbf{Y}) \equiv \hat{\theta} \in \Theta$. Avec les définitions du biais, la variance et l'erreur quadratique de la Sect. 2.8, on exprime cette dernière de la manière suivante :

$$\begin{aligned}
\mathbb{EQ}(\psi) &= \mathbb{E}\left\{(\theta - \psi(\mathbf{Y}))^2\right\} \\
&= \mathbb{E}\left\{[\theta - \mathbb{E}\{\psi(\mathbf{Y})\} + \mathbb{E}\{\psi(\mathbf{Y})\} - \psi(\mathbf{Y})]^2\right\} \\
&= \mathbb{E}\left\{([\theta - \mathbb{E}\{\psi(\mathbf{Y})\}] + [\mathbb{E}\{\psi(\mathbf{Y})\} - \psi(\mathbf{Y})])^2\right\} \\
&= \mathbb{E}\left\{[\theta - \mathbb{E}\{\psi(\mathbf{Y})\}]^2 + [\mathbb{E}\{\psi(\mathbf{Y})\} - \psi(\mathbf{Y})]^2 - 2[\theta - \mathbb{E}\{\psi(\mathbf{Y})\}][\mathbb{E}\{\psi(\mathbf{Y})\} - \psi(\mathbf{Y})]\right\} \\
&= \mathbb{E}\left\{[\theta - \mathbb{E}\{\psi(\mathbf{Y})\}]^2\right\} + \underbrace{\mathbb{E}\left\{[\mathbb{E}\{\psi(\mathbf{Y})\} - \psi(\mathbf{Y})]^2\right\}}_{=\mathbb{V}(\psi,\theta)} \\
&\qquad - 2\mathbb{E}\left\{[\theta - \mathbb{E}\{\psi(\mathbf{Y})\}][\mathbb{E}\{\psi(\mathbf{Y})\} - \psi(\mathbf{Y})]\right\} \\
&= [\underbrace{\theta - \mathbb{E}\{\psi(\mathbf{Y})\}}_{=\mathbb{B}(\psi,\theta)}]^2 + \mathbb{V}(\psi,\theta) - 2\mathbb{E}\left\{[\theta - \mathbb{E}\{\psi(\mathbf{Y})\}]\right\}\mathbb{E}\left\{[\mathbb{E}\{\psi(\mathbf{Y})\} - \psi(\mathbf{Y})]\right\} \\
&= \mathbb{B}(\psi,\theta)^2 + \mathbb{V}(\psi,\theta) - 2[\theta - \mathbb{E}\{\psi(\mathbf{Y})\}][\mathbb{E}\{\psi(\mathbf{Y})\} - \mathbb{E}\{\psi(\mathbf{Y})\}] \\
&= \mathbb{B}(\psi,\theta)^2 + \mathbb{V}(\psi,\theta)
\end{aligned}$$

(Remarque : nous avons délibérément omis l'indice $\mathbf{Y} \mid \theta$ pour $\mathbb{B}$, $\mathbb{V}$, $\mathbb{EQ}$ pour des questions de clarté de calculs.)

L'erreur quadratique $\mathbb{EQ}_{\mathbf{Y}|\theta}$ se décompose donc en une somme du carré du biais $\mathbb{B}_{\mathbf{Y}|\theta}$ et de la variance $\mathbb{V}_{\mathbf{Y}|\theta}$ de l'estimateur.

Pour l'erreur quadratique moyenne, nous obtenons un résultat semblable : $\mathbb{EQ}_{\Theta,\mathbf{Y}} = \mathbb{E}_\Theta(\mathbb{B}_{\mathbf{Y}|\Theta}(\psi,\Theta)^2) + \mathbb{V}_{,\Theta,\mathbf{Y}}(\psi)$. Cependant, l'espérance du carré du biais n'égalant pas le carré de l'espérance du biais, la formule semble moins « parfaite ». Elle peut être résumée

par « l'erreur quadratique moyenne est la somme de la moyenne des carrés des biais et de la variance moyenne ».

A.3 Coefficients de la droite de régression

Soit (Θ, Y) un couple de variables aléatoires continues, distribué sous la distribution $\mathrm{p}_{\Theta,Y}(\theta, y)$. quadratique sous la distribution jointe par rapport à la fonction affine $f(\Theta) = \alpha + \beta\Theta$, soit aux paramètres de régression α et β. Cette espérance s'écrit

$$\mathbb{E}_{\Theta,Y}(\|\psi(Y) - f(\Theta)\|^2) = \mathbb{E}_{\Theta,Y}(\|\psi(Y) - \alpha - \beta\Theta\|^2) = \mathbb{E}_{\Theta,Y}((\psi(Y) - \alpha - \beta\Theta)^2),$$

et ainsi, on a

$$\begin{aligned}
&\underset{f\in\mathcal{E}_{\text{fcts affines}}}{\operatorname{argmin}}\ \mathbb{E}_{\Theta,Y}(\|\psi(Y) - f(\Theta)\|^2) \\
&\quad= \underset{\alpha,\beta}{\operatorname{argmin}}\, \mathbb{E}_{\Theta,Y}((\psi(Y) - \alpha - \beta\Theta)^2) \\
&\quad= \underset{\alpha,\beta}{\operatorname{argmin}} \int\limits_{\Omega(\Theta\times Y)} (\psi(y) - \alpha - \beta\theta)^2\, \mathrm{p}_{\Theta,Y}(\theta, y)\, \mathrm{d}(\theta, y) \\
&\quad= \underset{\alpha,\beta}{\operatorname{argmin}} \int\limits_{\Omega(\Theta\times Y)} (\psi(y) - \alpha - \beta\theta)^2\, \mathrm{p}_{\Theta,Y}(\theta, y)\, \mathrm{d}(\theta, y) \\
&\quad= \underset{\alpha,\beta}{\operatorname{argmin}} \int\limits_{\Omega(\Theta\times Y)} \psi(y)^2 + \alpha^2 + \beta^2\theta^2 - 2\alpha\psi(y) - 2\beta\psi(y)\theta + 2\alpha\beta\theta\, \mathrm{p}_{\Theta,Y}(\theta, y)\, \mathrm{d}(\theta, y) \\
&\quad= \int\limits_{\Omega(Y)} \psi(y)^2\, \mathrm{p}_Y(y)\, \mathrm{d}y \\
&\qquad + \underset{\alpha,\beta}{\operatorname{argmin}} \Bigg[\alpha^2 + \beta^2 \int\limits_{\Omega(\Theta)} \theta^2\, \mathrm{p}_\Theta(\theta)\, \mathrm{d}\theta \\
&\qquad - 2\cdot\Bigg(\alpha \int\limits_{\Omega(Y)} \psi(y)\, \mathrm{p}_Y(y)\, \mathrm{d}y + \beta \int\limits_{\Omega(\Theta\times Y)} \theta\cdot\psi(y)\, \mathrm{p}_{\Theta,Y}(\theta, y)\, \mathrm{d}(\theta, y) \\
&\qquad - \alpha\beta \int\limits_{\Omega(\Theta)} \theta\, \mathrm{p}_\Theta(\theta)\, \mathrm{d}\theta\Bigg)\Bigg] \\
&\quad= \mathbb{E}_Y\left(\psi(Y)^2\right) \\
&\qquad + \underset{\alpha,\beta}{\operatorname{argmin}} \underbrace{\left[\alpha^2 + \beta^2\mathbb{E}_\Theta(\Theta^2) - 2\left(\alpha\mathbb{E}_Y(\psi(Y)) + \beta\mathbb{E}_{\Theta,Y}(\Theta\cdot\psi(Y)) - \alpha\beta\mathbb{E}_\Theta(\Theta)\right)\right]}_{=:Q(\alpha,\beta)}.
\end{aligned} \tag{A.4}$$

Pour trouver les paramètres $(\bar{\alpha}, \bar{\beta})$ qui minimisent l'espérance de l'erreur quadratique, il suffit de minimiser la fonction $Q(\alpha, \beta)$ par rapport au vecteur $[\alpha, \beta]^T$. Pour cela, on calcule le gradient de la fonction

$$\operatorname{grad} Q(\alpha, \beta) = 2\left[\alpha - \mathbb{E}_Y(\psi(Y)) + \beta\mathbb{E}_\Theta(\Theta);\quad \beta\mathbb{E}_\Theta(\Theta^2) - \mathbb{E}_{\Theta,Y}(\Theta\cdot\psi(Y)) + \alpha\mathbb{E}_\Theta(\theta)\right]^T. \tag{A.5}$$

La recherche du maximum se traduit par la recherche des points pour lesquels le gradient égale le vecteur nul :

$$\operatorname{grad} Q(\bar{\alpha}, \bar{\beta}) = [0, 0]^T \tag{A.6a}$$

$$\Leftrightarrow \quad \begin{cases} \bar{\alpha} - \mathbb{E}_Y(\psi(Y)) + \bar{\beta}\mathbb{E}_\Theta(\Theta) & = & 0 \\ \bar{\beta}\mathbb{E}_\Theta(\Theta^2) - \mathbb{E}_{\Theta,Y}(x \cdot \psi(Y)) + \bar{\alpha}\mathbb{E}_\Theta(\Theta) & = & 0 \end{cases} \tag{A.6b}$$

$$\Leftrightarrow \begin{cases} \boxed{\begin{array}{rcl} \bar{\alpha} & = & \mathbb{E}_Y(\psi(Y)) - \bar{\beta}\mathbb{E}_\Theta(\Theta) \\ \bar{\beta} & = & \dfrac{\mathbb{E}_{\Theta,Y}(\Theta \cdot \psi(Y)) - \bar{\alpha}\mathbb{E}_\Theta(\Theta)}{\mathbb{E}_\Theta(\Theta^2)} \end{array}} \end{cases} \tag{A.6c}$$

En substituant $\bar{\alpha}$ dans l'expression de $\bar{\beta}$ par son expression algébrique, on trouve pour

$$\bar{\beta} = \frac{\mathbb{E}_{\Theta,Y}(\Theta \cdot \psi(Y)) - \mathbb{E}_\Theta(\Theta)\mathbb{E}_Y(\psi(Y))}{\mathbb{E}_\Theta(\Theta^2) - \mathbb{E}_\Theta(\Theta)^2} = \frac{\mathbb{C}_{\Theta,Y}(\Theta, \psi(Y))}{\mathbb{V}_\Theta(\Theta)} = \frac{\text{covariance}(\Theta, \psi(Y))}{\text{variance}(\Theta)}. \tag{A.7}$$

Ainsi, nous avons pu relier les paramètres de la régression linéaire simple

1. aux espérances de Θ et de $\psi(Y)$ en ce qui concerne l'intercept α de la droite de régression, et
2. à la covariance entre Θ et $\psi(Y)$ ainsi qu'à la variance de Θ en ce qui concerne le facteur de pente β de la droite de régression.

A.4 Critère d'écart quadratique moyen pour une régression optimale

Nous calculons la valeur du critère d'écart quadratique moyen par rapport à la droite de régression pour les coefficients de régression optimaux, noté $\bar{\alpha}$ et $\bar{\beta}$ suivant la Sect. 2.8. Nous rappelons l'expression de l'écart quadratique moyen, voir Éqn. (2.31) :

$$\mathbb{J}_{\theta,\mathbf{Y},\psi}(\alpha, \beta) = \mathbb{E}_{\Theta,\mathbf{Y}}((\psi(\mathbf{Y}) - f(\Theta))^2) = \mathbb{E}_{\Theta,\mathbf{Y}}((\psi(\mathbf{Y}) - \alpha - \beta\Theta)^2). \tag{2.31}$$

Ce dernier est minimisé par les coefficients de régression optimaux, Éqs. (2.32) et (2.33) :

$$\bar{\alpha} = \mathbb{E}_\mathbf{Y}(\psi(\mathbf{Y})) - \bar{\beta}\mathbb{E}_\Theta(\Theta). \tag{2.32}$$

$$\bar{\beta} = \frac{\mathbb{E}_{\Theta,\mathbf{Y}}(\Theta \cdot \psi(\mathbf{Y})) - \mathbb{E}_\Theta(\Theta)\mathbb{E}_\mathbf{Y}(\psi(\mathbf{Y}))}{\mathbb{E}_\Theta(\Theta^2) - \mathbb{E}_\Theta(\Theta)^2} = \frac{\mathbb{C}_{\Theta,\mathbf{Y}}(\Theta, \psi(\mathbf{Y}))}{\mathbb{V}_\Theta(\Theta)}. \tag{2.33}$$

La valeur minimale de l'écart quadratique moyen est alors donnée par

$$\begin{aligned}
\mathbb{J}_{\theta,\mathbf{Y},\psi}(\bar{\alpha},\bar{\beta}) &= \mathbb{E}_{\Theta,\mathbf{Y}}\left\{\left[\psi(\mathbf{Y})-\bar{\alpha}-\bar{\beta}\Theta\right]^2\right\} \\
&= \mathbb{E}_{\Theta,\mathbf{Y}}\left\{\left[\psi(\mathbf{Y})-\mathbb{E}_{\mathbf{Y}}(\psi(\mathbf{Y}))+\bar{\beta}\mathbb{E}_{\Theta}(\Theta)-\bar{\beta}\Theta\right]^2\right\} \\
&= \mathbb{E}_{\Theta,\mathbf{Y}}\left\{\left[(\psi(\mathbf{Y})-\mathbb{E}_{\mathbf{Y}}(\psi(\mathbf{Y})))-\left(\bar{\beta}\Theta-\bar{\beta}\mathbb{E}_{\Theta}(\Theta)\right)\right]^2\right\} \\
&= \mathbb{E}_{\Theta,\mathbf{Y}}\Big\{(\psi(\mathbf{Y})-\mathbb{E}_{\mathbf{Y}}(\psi(\mathbf{Y})))^2+\bar{\beta}^2\left(\Theta-\mathbb{E}_{\Theta}(\Theta)\right)^2 \\
&\qquad -2\left(\psi(\mathbf{Y})-\mathbb{E}_{\mathbf{Y}}(\psi(\mathbf{Y}))\right)\bar{\beta}\left(\Theta-\mathbb{E}_{\Theta}(\Theta)\right)\Big\} \\
&= \underbrace{\mathbb{E}_{\mathbf{Y}}\left\{(\psi(\mathbf{Y})-\mathbb{E}_{\mathbf{Y}}(\psi(\mathbf{Y})))^2\right\}}_{=\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))}+\bar{\beta}^2\underbrace{\mathbb{E}_{\Theta}\left\{(\Theta-\mathbb{E}_{\Theta}(\Theta))^2\right\}}_{=\mathbb{V}_{\Theta}(\Theta)} \\
&\qquad -2\bar{\beta}\underbrace{\mathbb{E}_{\mathbf{Y}}\left(\psi(\mathbf{Y})-\mathbb{E}_{\mathbf{Y}}(\psi(\mathbf{Y}))\right)}_{=0}\underbrace{\mathbb{E}_{\Theta}\left(\Theta-\mathbb{E}_{\Theta}(\Theta)\right)}_{=0} \\
&\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))-\bar{\beta}^2\mathbb{V}_{\Theta}(\Theta) \\
&\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))-\frac{\mathbb{C}_{\Theta,\mathbf{Y}}(\Theta,\psi(\mathbf{Y}))^2}{\mathbb{V}_{\Theta}(\Theta)^2}\mathbb{V}_{\Theta}(\Theta). \qquad (*)
\end{aligned}$$

En utilisant la définition du coefficient de détermination, Éqn. (2.34), on obtient

$$\begin{aligned}
R^2_{\bar{\alpha},\bar{\beta}} &= 1-\frac{\mathbb{J}_{\theta,\mathbf{Y},\psi}(\bar{\alpha},\bar{\beta})}{\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))} \\
&\overset{(*)}{=} 1-\frac{\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))-\frac{\mathbb{C}_{\Theta,\mathbf{Y}}(\Theta,\psi(\mathbf{Y}))^2}{\mathbb{V}_{\Theta}(\Theta)^2}\mathbb{V}_{\Theta}(\Theta)}{\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))} \\
&= 1-\frac{\mathbb{V}_{\Theta}(\Theta)\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))-\mathbb{C}_{\Theta,\mathbf{Y}}(\Theta,\psi(\mathbf{Y}))^2}{\mathbb{V}_{\Theta}(\Theta)\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))} \\
&= \frac{\mathbb{C}_{\Theta,\mathbf{Y}}(\Theta,\psi(\mathbf{Y}))^2}{\mathbb{V}_{\Theta}(\Theta)\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))}. \qquad (**)
\end{aligned}$$

En utilisant cette identité, nous poursuivons les calculs :

$$\begin{aligned}
\mathbb{J}_{\theta,\mathbf{Y},\psi}(\bar{\alpha},\bar{\beta}) &= \mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))-\frac{\mathbb{C}_{\Theta,\mathbf{Y}}(\Theta,\psi(\mathbf{Y}))^2}{\mathbb{V}_{\Theta}(\Theta)^2}\mathbb{V}_{\Theta}(\Theta) \\
&= \mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))-\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))\underbrace{\frac{\mathbb{C}_{\Theta,\mathbf{Y}}(\Theta,\psi(\mathbf{Y}))^2}{\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))\mathbb{V}_{\Theta}(\Theta)}}_{=R^2_{\bar{\alpha},\bar{\beta}}\text{ d'après }(**)} \\
&= \mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))(1-R^2_{\bar{\alpha},\bar{\beta}}).
\end{aligned}$$

Ainsi, nous avons obtenu, en plus de l'expression explicite du coefficient de détermination dans un cas de régression simple optimale, la valeur de la borne inférieure dudit coefficient :

$$\mathbb{J}_{\theta,\mathbf{Y},\psi}(\alpha,\beta) \geq \mathbb{J}_{\theta,\mathbf{Y},\psi}(\bar{\alpha},\bar{\beta}) = \mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))(1-R^2_{\bar{\alpha},\bar{\beta}}). \qquad (A.8)$$

A.5 Preuve de la Proposition ⟨2.29⟩

Rappel du contexte

Dans la Sect. 2.8, nous avons introduit les mesures d'erreurs usuelles (biais, variance, erreur quadratique) et la droite de régression entre des valeurs explicatives Θ et des va-

leurs estimées $\psi(\mathbf{Y})$.

Nous cherchons un lien entre l'estimateur parfait (Déf. ⟨2.26⟩) et la droite de régression qui est une bissectrice expliquant 100 % de la variation des estimations linéairement par la variable explicative.

Proposition *Soient Θ une variable aléatoire explicative et $\mathbf{Y}$ une variable aléatoire régressée. Soit ensuite $\psi : \Omega(\mathbf{Y}) \to \Omega(\Theta)$, $\mathbf{Y} \mapsto \psi(\mathbf{Y})$ un estimateur.*

L'estimateur ψ est parfait si et seulement si la droite de régression $f(\theta) = \alpha + \beta\theta$ est la bissectrice et que le coefficient de détermination vaut 1, i.e. les coefficients de régression valent $(\alpha, \beta, R^2) = (0, 1, 1)$.

Preuve

"⇒" Soit ψ un estimateur parfait. Alors, l'erreur quadratique moyenne s'annule, $\mathbb{E}_{\Theta,\mathbf{Y}}(\psi) = \mathbb{E}_{\Theta}\left(\mathbb{E}_{\mathbf{Y}|\Theta}\left([\psi(\mathbf{Y}) - \Theta]^2\right)\right) = 0$. Comme l'espérance est prise sur des valeurs positives ou nulles (car il s'agit d'un carré), on a immédiatement que $\psi(\mathbf{Y}) = \theta = 0 + 1 \cdot \theta = f(\theta)$ avec donc $\alpha = 0$ et $\beta = 1$.

Ainsi, on a aussi pour la covariance l'égalité avec la variance de Θ : $\mathbb{C}_{\Theta,\mathbf{Y}}(\Theta, \psi(\mathbf{Y})) = \mathbb{C}_{\Theta,\Theta}(\Theta, \Theta) = \mathbb{V}_{\Theta}(\Theta)$. Avec la définition de l'Éqn. (2.35), on a pour le coefficient de détermination $R^2 = \mathbb{C}_{\Theta,\mathbf{Y}}(\Theta, \psi(\mathbf{Y}))/(\mathbb{V}_{\Theta}(\Theta)\mathbb{V}_{\mathbf{Y}}(\psi(\mathbf{Y}))) = \mathbb{V}_{\Theta}(\Theta)^2/\mathbb{V}_{\Theta}(\Theta)^2 = 1$.

"⇐" Soit maintenant la droite de régression déterminée par les coefficients $(\alpha, \beta, R^2) = (0, 1, 1)$. De la définition de α, Éqn. (2.32), on déduit $\mathbb{E}_{\mathbf{Y}}(\psi(\mathbf{Y})) - \beta\mathbb{E}_{\Theta}(\Theta) = 0$. Avec $\beta = 1$, on obtient alors

$$\mathbb{E}_{\mathbf{Y}}(\psi(\mathbf{Y})) = \mathbb{E}_{\Theta}(\Theta), \tag{A.9}$$

l'égalité des espérances, *i.e.* des premiers moments. On en déduit également

$$\mathbb{E}_{\Theta,\mathbf{Y}}\left(\Theta - \psi(\mathbf{Y})\right) = 0. \tag{A.10}$$

Le biais moyen est nul.

En considérant la définition de β, Éqn. (2.33), et que sa valeur vaut 1, on déduit l'égalité entre la covariance de Θ et $\psi(\mathbf{Y})$ et la variance de Θ. En développant les expressions et en utilisant Éqn. (A.9), on obtient

$$\mathbb{E}_{\Theta,\mathbf{Y}}(\Theta \cdot \psi(\mathbf{Y})) = \mathbb{E}_{\Theta}(\Theta^2), \tag{A.11}$$

i.e. le premier moment de la loi jointe pour Θ et $\psi(\mathbf{Y})$ est égale au deuxième moment de la loi pour Θ.

Ensuite, par la définition du coefficient de détermination R^2, Éqn. (2.35), et de sa valeur unitaire, la covariance de Θ et $\psi(\mathbf{Y})$ égale le produit des variances pour Θ et pour $\psi(\mathbf{Y})$. Avec ce qui précède, on déduit que la variance de Θ au carré égale le produit des variances, ce qui amène l'égalité entre la variance de Θ et celle de $\psi(\mathbf{Y})$. En développant les expressions des variances et en utilisant Éqn. (A.9), on trouve l'égalité des deuxièmes moments des lois, *i.e.*

$$\mathbb{E}_{\Theta}(\Theta^2) = \mathbb{E}_{\mathbf{Y}}(\psi(\mathbf{Y})^2). \tag{A.12}$$

Avec tout ce qui précède, on exprime l'erreur quadratique moyenne

$$\begin{aligned}
\mathbb{EQ}_{\Theta,\mathbf{Y}}(\psi) &= \mathbb{E}_{\Theta,\mathbf{Y}}\left((\psi(\mathbf{Y}) - \Theta)^2\right) \\
&= \underbrace{\mathbb{E}_{\mathbf{Y}}\left(\psi(\mathbf{Y})^2\right)}_{\text{Éqn. (A.12)}} - 2\mathbb{E}_{\Theta,\mathbf{Y}}(\Theta \cdot \psi(\mathbf{Y})) + \mathbb{E}_{\Theta}(\Theta^2) \\
&= 2\left[\mathbb{E}_{\Theta}(\Theta^2) - \underbrace{\mathbb{E}_{\Theta,\mathbf{Y}}(\Theta \cdot \psi(\mathbf{Y}))}_{\text{Éqn. (A.11)}}\right] \\
&= 0
\end{aligned}$$

qui s'annule. L'estimateur ψ est donc parfait. ■

Ainsi, nous avons déduit un lien entre un estimateur parfait – dans le sens introduit précédemment – et les coefficients de la droite de régression.

A.6 Les matrices des systèmes

Nous allons exprimer l'Éqn. (3.11) d'une manière matricielle afin d'accentuer le caractère linéaire des paramètres $\boldsymbol{\kappa} = [\kappa_1, \ldots, \kappa_I]$, $\boldsymbol{\kappa}^* = [\kappa_1^*, \ldots, \kappa_I^*]$ et $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_I]$ par rapport aux données $\mathbf{Y}$. Pour ce faire, soit $\boldsymbol{y} \in \mathbb{R}^N$ le vecteur de dimension $N = N_{\mathscr{C}} \cdot N_{\mathscr{S}}$ résultant de la concaténation des colonnes de la matrice $\mathbf{Y} \in \mathbb{R}^{N_{\mathscr{C}} \times N_{\mathscr{S}}}$, et soit $\boldsymbol{b} \in \mathbb{R}^N$ la vectorisation de la matrice de bruit $\mathbf{B}$. Nous pouvons alors réécrire la sortie du système sous la forme

$$\boldsymbol{y} = \mathbf{H}\boldsymbol{\kappa} + \mathbf{H}^*\boldsymbol{\kappa}^* + \boldsymbol{b}, \qquad \mathbf{H}, \mathbf{H}^* \in \mathbb{R}^{N \times I} \tag{A.13}$$

et

$$\boldsymbol{y} = \mathbf{G}\boldsymbol{\xi} + \boldsymbol{b}, \qquad \mathbf{G} \in \mathbb{R}^{N \times I}. \tag{A.14}$$

Ceci montre la relation linéaire liant les quantités de peptides aux données, et celle entre le gain du système et l'observation. Pour définir ces matrices, posons la signature élémentaire du peptide natif i

$$\mathbf{F}_i = \sum_{j=1}^{J} \sum_{k=0}^{K} \pi_{ij} \pi'_{ijk} \mathscr{S}_{ijk} \mathscr{C}_i^{\mathrm{T}} \tag{A.15}$$

et celle du peptide i issu du marquage isotopique

$$\mathbf{F}_i^* = \sum_{j=1}^{J} \sum_{k=0}^{K} \pi_{ij} \pi'^*_{ijk} \mathscr{S}^*_{ijk} \mathscr{C}_i^{\mathrm{T}}. \tag{A.16}$$

La sortie modèle étant la somme des signatures, pondérées par le gain du peptide et par sa quantité, nous avons

$$\mathbf{M} = \sum_{i=1}^{I} \kappa_i \xi_i \mathbf{F}_i \qquad \text{et} \qquad \mathbf{M}^* = \sum_{i=1}^{I} \kappa_i^* \xi_i \mathbf{F}_i^*.$$

Soient $\mathbf{F}$ et $\mathbf{F}^*$ les matrices de dimension $N \times I$ qui s'obtiennent en juxtaposant horizontalement les vectorisations des matrices $\overrightarrow{\mathbf{F}_i}$ et $\overrightarrow{\mathbf{F}_i^*}$ respectivement :

$$\mathbf{F} = \begin{bmatrix} | & | & & | \\ \overrightarrow{\mathbf{F}_1} & \overrightarrow{\mathbf{F}_2} & \ldots & \overrightarrow{\mathbf{F}_I} \\ | & | & & | \end{bmatrix} \qquad \text{et} \qquad \mathbf{F}^* = \begin{bmatrix} | & | & & | \\ \overrightarrow{\mathbf{F}_1^*} & \overrightarrow{\mathbf{F}_2^*} & \ldots & \overrightarrow{\mathbf{F}_I^*} \\ | & | & & | \end{bmatrix}.$$

Vu sous cet angle, la vectorisation des matrices $\mathbf{M}$ et $\mathbf{M}^*$ peut s'écrire sous forme matricielle :

$$\boldsymbol{m} = \mathbf{F}(\boldsymbol{\kappa} \odot \boldsymbol{\xi}) \qquad \text{et} \qquad \boldsymbol{m}^* = \mathbf{F}^*(\boldsymbol{\kappa}^* \odot \boldsymbol{\xi})$$

où $\odot$ désigne la multiplication matricielle de Hadamard. Appliqué aux vecteurs, elle peut être exprimée de la manière suivante

$$\boldsymbol{a} \odot \boldsymbol{b} = \mathrm{diag}(\boldsymbol{a})\boldsymbol{b}.$$

La multiplication terme à terme est commutative, car

$$\boldsymbol{b} \odot \boldsymbol{a} = \mathrm{diag}(\boldsymbol{b})\boldsymbol{a} = [b_1 a_1, \ldots, b_N a_N] = [a_1 b_1, \ldots, a_N b_N] = \mathrm{diag}(\boldsymbol{a})\boldsymbol{b} = \boldsymbol{a} \odot \boldsymbol{b}.$$

Nous réécrivons la sortie du système de manière matricielle :

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{m} + \boldsymbol{m}^* + \boldsymbol{b} \\
&= \mathbf{F}(\boldsymbol{\kappa} \odot \boldsymbol{\xi}) + \mathbf{F}^*(\boldsymbol{\kappa}^* \odot \boldsymbol{\xi}) + \boldsymbol{b} \\
&= \underbrace{\left[\mathbf{F}\,\mathrm{diag}(\boldsymbol{\kappa}) + \mathbf{F}^*\,\mathrm{diag}(\boldsymbol{\kappa}^*)\right]}_{=:\mathbf{G}} \boldsymbol{\xi} + \boldsymbol{b} \\
&= \underbrace{\left[\mathbf{F}\,\mathrm{diag}(\boldsymbol{\xi})\right]}_{=:\mathbf{H}} \boldsymbol{\kappa} + \underbrace{\left[\mathbf{F}^*\,\mathrm{diag}(\boldsymbol{\xi})\right]}_{=:\mathbf{H}^*} \boldsymbol{\kappa}^*.
\end{aligned}$$

Définition *Les matrices du système* $\mathbf{H}, \mathbf{H}^*, \mathbf{G} \in \mathbb{R}^{N \times I}$ *avec* $N = N_{\mathscr{C}} \cdot N_{\mathscr{S}}$ *sont définies à partir des matrices des signatures élémentaires* $\mathbf{F}$ *et* $\mathbf{F}^*$ *et des paramètres* $\boldsymbol{\kappa}$, $\boldsymbol{\kappa}^*$, $\boldsymbol{\xi}$ *de la manière suivante :*

$$\mathbf{H} = \mathbf{F}\,\mathrm{diag}(\boldsymbol{\xi}), \tag{A.17a}$$

$$\mathbf{H}^* = \mathbf{F}^*\,\mathrm{diag}(\boldsymbol{\xi}), \tag{A.17b}$$

$$\mathbf{G} = \mathbf{F}\,\mathrm{diag}(\boldsymbol{\kappa}) + \mathbf{F}^*\,\mathrm{diag}(\boldsymbol{\kappa}^*). \tag{A.17c}$$

B Calculs de lois

Dans ce chapitre annexe, nous déduirons les lois *a posteriori* conditionnelles pour les paramètres dont la loi *a priori* est conjuguée par la vraisemblance conditionnelle. Nous allons d'abord exprimer les lois *a posteriori* conditionnelles qui interviennent dans la quantification. Comme dans ce que nous avons présenté dans le corps du document, les paramètres techniques sont les mêmes pour les trois étapes, les calculs de leurs lois *a posteriori* conditionnelles ne sont pas à refaire pour l'apprentissage (modulo l'indice de l'individu) et la classification. Il suffit donc de présenter les calculs des lois *a posteriori* conditionnelles pour le paramètre biologique, et également pour les paramètres classe (apprentissage) ou le paramètre individu (état biologique, classification).

B.1 Conjugaison d'une loi normale par une vraisemblance normale

Nous traitons souvent une loi *a priori* normale conjuguée par une vraisemblance normale. Pour ne pas mener plusieurs fois le même calcul à l'appellation du paramètre près, nous déduisons le résultat du produit de ces deux lois dans un cadre plus général. Pour pouvoir appliquer les calculs suivants, nous notons [1]

- le paramètre d'intérêt $\boldsymbol{x} \in \mathbb{R}^N$,
- sa loi *a priori* $\mathrm{p}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_x, \boldsymbol{\Gamma}_x)$,
- sa fonction de vraisemblance $\mathrm{p}(\mathbf{Y} \mid \boldsymbol{x}) = \mathcal{N}(\mathbf{Y}; \mathbf{H}\boldsymbol{x}, \boldsymbol{\Gamma}_\mathbf{Y})$.

Pour la loi *a posteriori*, nous avons

$$\begin{aligned}
\mathrm{p}(\boldsymbol{x} \mid \mathbf{Y}) &\propto \mathrm{p}(\boldsymbol{x}) \cdot \mathrm{p}(\mathbf{Y} \mid \boldsymbol{x}) \\
&= \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_x, \boldsymbol{\Gamma}_x) \cdot \mathcal{N}(\mathbf{Y}; \mathbf{H}\boldsymbol{x}, \boldsymbol{\Gamma}_\mathbf{Y}) \\
&\propto \exp\Big(-\frac{1}{2}\Big[\underbrace{(\boldsymbol{x} - \boldsymbol{m}_x)^\mathrm{T} \boldsymbol{\Gamma}_x (\boldsymbol{x} - \boldsymbol{m}_x) + (\mathbf{Y} - \mathbf{H}\boldsymbol{x})^\mathrm{T} \boldsymbol{\Gamma}_\mathbf{Y} (\mathbf{Y} - \mathbf{H}\boldsymbol{x})}_{=:Q(\boldsymbol{x})} \Big]\Big)
\end{aligned}$$

Nous allons montrer que nous pouvons ramener l'expression de $Q(\boldsymbol{x})$ à l'expression

$$\tilde{Q}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{m}_x^{\mathrm{post}})^\mathrm{T} \boldsymbol{\Gamma}_x^{\mathrm{post}} (\boldsymbol{x} - \boldsymbol{m}_x^{\mathrm{post}}).$$

Ainsi, nous aurons déduit la loi normale $\mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_x^{\mathrm{post}}, \boldsymbol{\Gamma}_x^{\mathrm{post}})$. Pour cela, nous proposons soit de développer $Q(\boldsymbol{x})$ et $\tilde{Q}(\boldsymbol{x})$, soit d'identifier les moments de $Q(\boldsymbol{x})$ et de $\tilde{Q}(\boldsymbol{x})$.

B.1.1 Identification des facteurs

Développons les expressions :

$$\begin{aligned}
Q(\boldsymbol{x}) &= (\boldsymbol{x} - \boldsymbol{m}_x)^\mathrm{T} \boldsymbol{\Gamma}_x (\boldsymbol{x} - \boldsymbol{m}_x) + (\mathbf{Y} - \mathbf{H}\boldsymbol{x})^\mathrm{T} \boldsymbol{\Gamma}_\mathbf{Y} (\mathbf{Y} - \mathbf{H}\boldsymbol{x}) \\
&= \boldsymbol{x}^\mathrm{T}\boldsymbol{\Gamma}_x\boldsymbol{x} - 2\boldsymbol{x}^\mathrm{T}\boldsymbol{\Gamma}_x\boldsymbol{m}_x + \boldsymbol{m}_x^\mathrm{T}\boldsymbol{\Gamma}_x\boldsymbol{m}_x + \mathbf{Y}^\mathrm{T}\boldsymbol{\Gamma}_\mathbf{Y}\mathbf{Y} - 2(\mathbf{H}\boldsymbol{x})^\mathrm{T}\boldsymbol{\Gamma}_\mathbf{Y}\mathbf{Y} + (\mathbf{H}\boldsymbol{x})^\mathrm{T}\boldsymbol{\Gamma}_\mathbf{Y}(\mathbf{H}\boldsymbol{x}) \\
&= \boldsymbol{x}^\mathrm{T}(\boldsymbol{\Gamma}_x + \mathbf{H}^\mathrm{T}\boldsymbol{\Gamma}_\mathbf{Y}\mathbf{H})\boldsymbol{x} - 2\boldsymbol{x}^\mathrm{T}(\boldsymbol{\Gamma}_x\boldsymbol{m}_x + \mathbf{H}^\mathrm{T}\boldsymbol{\Gamma}_\mathbf{Y}\mathbf{Y}) + \boldsymbol{m}_x^\mathrm{T}\boldsymbol{\Gamma}_x\boldsymbol{m}_x + \mathbf{Y}^\mathrm{T}\boldsymbol{\Gamma}_\mathbf{Y}\mathbf{Y}
\end{aligned}$$

1. indépendamment des notations du document

et

$$\begin{aligned}\tilde{Q}(\boldsymbol{x}) &= (\boldsymbol{x} - \boldsymbol{m}_x^{\text{post}})^{\mathrm{T}} \boldsymbol{\Gamma}_x^{\text{post}} (\boldsymbol{x} - \boldsymbol{m}_x^{\text{post}}) \\ &= \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Gamma}_x^{\text{post}} \boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Gamma}_x^{\text{post}} \boldsymbol{m}_x^{\text{post}} + \boldsymbol{m}_x^{\text{post}} \boldsymbol{\Gamma}_x^{\text{post}} \boldsymbol{m}_x^{\text{post}}\end{aligned}$$

En faisant une identification terme à terme, on constate que

1. $\boldsymbol{\Gamma}_x^{\text{post}} = (\boldsymbol{\Gamma}_x + \mathbf{H}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{H})$ et
2. $\boldsymbol{\Gamma}_x^{\text{post}} \boldsymbol{m}_x^{\text{post}} = (\boldsymbol{\Gamma}_x \boldsymbol{m}_x + \mathbf{H}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{Y})$, d'où $\boldsymbol{m}_x^{\text{post}} = (\boldsymbol{\Gamma}_x^{\text{post}})^{-1} (\boldsymbol{\Gamma}_x \boldsymbol{m}_x + \mathbf{H}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{Y})$

La moyenne *a posteriori* $\boldsymbol{m}_x^{\text{post}}$ est donc une moyenne pondérée entre la moyenne de loi *a priori* et la valeur déduite des données. La précision *a posteriori*, elle, est une somme de précision : à la précision *a priori* s'ajoute la précision des données mise à l'échelle du paramètre $\boldsymbol{x}$.

B.1.2 Identification des moments

Nous allons obtenir maintenant le même résultat, mais avec un moyen différent. Une loi normale est définie par la moyenne et la précision, *i.e.* son premier et l'inverse de son deuxième moment. Tous les deux se déduisent de l'argument de l'exponentielle $Q(\boldsymbol{x})$: le premier moment $\boldsymbol{m}_x^{\text{post}}$ est la valeur telle que $\frac{\partial}{\partial \boldsymbol{x}} Q(\boldsymbol{x}) = 0$, le deuxième $\boldsymbol{\Gamma}_x^{\text{post}}$ vaut $\frac{1}{2} \frac{\partial^2}{\partial \boldsymbol{x}^2} Q(\boldsymbol{x})$.

En profitant des calculs précédents, les deux dérivés se calculent facilement. On trouve

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{x}} Q(\boldsymbol{x}) &= 2(\boldsymbol{\Gamma}_x + \mathbf{H}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{H}) \boldsymbol{x} - 2(\boldsymbol{\Gamma}_x \boldsymbol{m}_x + \mathbf{H}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{Y}) \\ &\overset{!}{=} 0 \\ &\Leftrightarrow \\ (\boldsymbol{\Gamma}_x + \mathbf{H}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{H}) \boldsymbol{x}_0 &= \boldsymbol{\Gamma}_x \boldsymbol{m}_x + \mathbf{H}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{Y} \\ \boldsymbol{x}_0 &= (\boldsymbol{\Gamma}_x + \mathbf{H}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{H})^{-1} (\boldsymbol{\Gamma}_x \boldsymbol{m}_x + \mathbf{H}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{Y}) = \boldsymbol{m}_x^{\text{post}}\end{aligned}$$

et pour l'inverse du deuxième moment

$$\frac{1}{2} \frac{\partial^2}{\partial \boldsymbol{x}^2} Q(\boldsymbol{x}) = \boldsymbol{\Gamma}_x + \mathbf{H}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{H} = \boldsymbol{\Gamma}_x^{\text{post}}.$$

Il s'agit (heureusement !) du même résultat qu'avec l'identification directe terme à terme.

Résumons le résultat de cette section :

$$\boxed{\boldsymbol{\Gamma}_x^{\text{post}} = (\boldsymbol{\Gamma}_x + \mathbf{H}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{H})} \quad \text{et} \quad \boxed{\boldsymbol{m}_x^{\text{post}} = (\boldsymbol{\Gamma}_x^{\text{post}})^{-1} (\boldsymbol{\Gamma}_x \boldsymbol{m}_x + \mathbf{H}^{\mathrm{T}} \boldsymbol{\Gamma}_{\mathbf{Y}} \mathbf{Y})}. \tag{B.1}$$

B.1.3 Cas extrêmes

Regardons à présent quelques cas extrêmes de la loi *a priori* et de la vraisemblance.

1. Que se passe-t-il si la loi *a priori* conditionnelle pour $\boldsymbol{x}$ est non informative [2] ? C'est en effet la situation la plus courante dans les analyses bayésiennes. La précision tend vers 0, la loi ressemble de plus en plus à une loi uniforme sur tout $\mathbb{R}$. Au vu de l'Éqn. (B.1), les paramètres *a posteriori* sont déterminés par les connaissances

2. Par loi non informative, nous entendons toute loi dans laquelle aucune information subjective a été injectée. C'est la cas notamment des lois issues du principe de Jeffreys [3, Sect. 2.9], voir Sect. 2.4. Dans ce contexte, la loi normale avec une précision nulle est identique à la loi uniforme sur tout l'ensemble $\mathbb{R}$ et donc non informative.

qu'apportent les données : $\mathbf{\Gamma}_x^{\text{post}} \to \mathbf{H}^{\mathrm{T}}\mathbf{\Gamma}_\mathbf{Y}\mathbf{H}$ et $\boldsymbol{m}_x^{\text{post}} \to (\mathbf{H}^{\mathrm{T}}\mathbf{\Gamma}_\mathbf{Y}\mathbf{H})^{-1}(\mathbf{H}^{\mathrm{T}}\mathbf{\Gamma}_\mathbf{Y}\mathbf{Y}) = (\mathbf{H}^{\mathrm{T}}\mathbf{\Gamma}_\mathbf{Y}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{\Gamma}_\mathbf{Y}\mathbf{Y}$ qui est la solution au sens des moindres carrés généralisés. Si $\mathbf{\Gamma}_\mathbf{Y}$ est diagonale, elle devient la solution au sens des moindres carrés.

2. Si les données observées sont très bruitées, c'est-à-dire $\mathbf{\Gamma}_\mathbf{Y}$ est proche de 0, alors sont influence par rapport aux autres variables dans les calculs est négigeable. La précision *a posteriori* est donc principalement composée de l'*a priori*, et très peu des déductions à partir des données. La loi *a posteriori* est donc très proche de la loi *a priori*.
3. Nous nous trouvons dans un scénario similaire quand la loi *a priori* est très piquée, très concentrée autour d'un point. Dans ce cas, la valeur de la précision est très grande, beaucoup plus grande que la précision des données qui devient négligeable. Dans ce cas, c'est également la loi *a priori* qui a plus de poids dans le calcul de la loi *a posteriori*.
4. Finalement, si les perturbations des mesures sont inexistantes, donc $\mathbf{\Gamma}_\mathbf{Y}$ tend vers l'infini, l'apport de la loi *a priori* est négigeable. Toute confiance sera pour l'information apportée par les données.

B.2 Conjugaison d'une loi gamma par une vraisemblance normale

Pour chaque point de mesure n, nous modélisons un bruit blanc gaussien, centré, de moyenne nulle et d'inverse variance γ, *i.e.* $\varepsilon_n \sim \mathcal{N}(0, \gamma)$. Le spectre a N points de mesure, nous avons autant de réalisations de la même distribution du bruit. Notons dans cette section la sortie instrument $\mathbf{Y}$, la sortie modèle $\mathbf{M}(\theta)$ attachée aux paramètres de génération de données θ.

La loi *a priori* pour le paramètre de bruit γ est une loi gamma que nous écrivons sous la forme suivante :

$$\mathrm{p}(\gamma) = \mathcal{G}(\gamma; \alpha, \beta) = \frac{\gamma^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} \exp\left(-\gamma/\beta\right) \mathbb{1}_{\mathbb{R}_+} \tag{B.2}$$

où $\alpha > 0$ définit la forme (*shape*) de la fonction, $\beta > 0$ l'échelle (*scale*), correspondant à la définition donnée dans [38, App. 1].

Remarque *La moyenne de la loi gamma est donnée par le produit des deux hyperparamètres, $\alpha\beta$, le mode par $(\alpha - 1)\beta$, et la variance par $\alpha\beta^2$.*

La loi *a posteriori* conditionnelle pour γ étant proportionnelle au produit « *a priori* $\times$ vraisemblance », nous avons

$$\begin{aligned}
\mathrm{p}(\gamma \mid \mathbf{Y}, \boldsymbol{\kappa}, \boldsymbol{\kappa}^*, \boldsymbol{\xi}, \boldsymbol{\tau}) &\propto \mathrm{p}(\gamma) \cdot \mathrm{p}(\mathbf{Y} \mid \boldsymbol{\kappa}, \boldsymbol{\kappa}^*, \boldsymbol{\xi}, \boldsymbol{\tau}, \gamma) \\
&= \mathcal{G}(\gamma; \alpha, \beta) \cdot \mathcal{N}(\mathbf{Y}; \mathbf{M}(\theta), \gamma) \\
&\propto \gamma^{\alpha-1}\gamma^{N/2} \exp\left(-\gamma/\beta\right) \exp\left(-\frac{\gamma}{2} \left\|\mathbf{Y} - \mathbf{M}(\theta)\right\|^2\right) \\
&= \gamma^{\alpha+N/2-1} \exp\left(-\gamma\left(\beta^{-1} + \left\|\mathbf{Y} - \mathbf{M}(\theta)\right\|^2/2\right)\right) \\
&\propto \mathcal{G}(\gamma, \alpha^{\text{post}}, \beta^{\text{post}}) \\
&\qquad \text{avec } \alpha^{\text{post}} = \alpha + N/2 \text{ et } \beta^{\text{post}} = \left(\beta^{-1} + \left\|\mathbf{Y} - \mathbf{M}(\theta)\right\|^2/2\right)^{-1}.
\end{aligned} \tag{B.3}$$

Pour introduire le moins d'information *a priori* possible, nous faisons tendre les paramètres *a priori* $\alpha \to 0$ et $\beta \to \infty$. La loi tend ainsi vers la loi de Jeffreys $\mathrm{p}(\gamma) \propto \gamma^{-1}$. Ce faisant, la loi *a posteriori* conditionnelle – qui est toujours une loi gamma – a les paramètres $\alpha^{\text{post}} = N/2$ et $\beta^{\text{post}} = 2/\left\|\mathbf{Y} - \mathbf{M}(\theta)\right\|^2$.

B.3 Conjugaison d'une loi normale-wishartienne par une vraisemblance normale

B.3.1 Loi normale-wishartienne

Nous voulons décrire la loi pour le couple $(\boldsymbol{m}, \boldsymbol{\Gamma})$, et nous exprimons sa loi jointe $\mathrm{p}(\boldsymbol{m}, \boldsymbol{\Gamma}) = \mathrm{p}(\boldsymbol{\Gamma})\,\mathrm{p}(\boldsymbol{m} \mid \boldsymbol{\Gamma})$. Pour la matrice de précision, nous avons recours à la loi de Wishart (ou loi wishartienne) de dimension P, $\mathcal{W}_P(\boldsymbol{\Gamma}; \boldsymbol{\Lambda}, \nu)$ de matrice d'échelle $\boldsymbol{\Lambda}$ et de ν degrés de liberté. Pour le paramètre de la moyenne, $\boldsymbol{m}$, on choisit une loi multinormale de moyenne $\boldsymbol{\mu}$ et de matrice de précision $\eta\boldsymbol{\Gamma}$, $\mathcal{N}_P(\boldsymbol{m}; \boldsymbol{\mu}, \eta\boldsymbol{\Gamma})$. Résumons :

$$\begin{aligned}
\boldsymbol{\Gamma} \sim \mathcal{W}_P(\boldsymbol{\Lambda}, \nu) &= \frac{|\boldsymbol{\Lambda}|^{-\nu/2}}{2^{\nu P/2}(2\pi)^P \Gamma_P(\nu/2)} \left|\boldsymbol{\Gamma}\right|^{(\nu-P-1)/2} \exp\left(-\tfrac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma})\right) \\
\boldsymbol{m} \sim \mathcal{N}_P(\boldsymbol{\mu}, \eta\boldsymbol{\Gamma}) &= (2\pi)^{-P/2} \left|\boldsymbol{\Gamma}\right|^{1/2} \exp\left(-\tfrac{\eta}{2}(\boldsymbol{m}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{m}-\boldsymbol{\mu})\right)
\end{aligned}$$

où $\Gamma_P(\cdot)$ est la fonction Gamma P-dimensionnelle.

Le produit de ces deux lois forme la loi *a priori* du couple $(\boldsymbol{m}, \boldsymbol{\Gamma})$. Cette loi est connue sous le nom de *loi normale-wishartienne* :

$$\begin{aligned}
&\mathrm{p}(\boldsymbol{m}, \boldsymbol{\Gamma}) = \mathcal{NW}_P(\boldsymbol{m}, \boldsymbol{\Gamma}; \boldsymbol{\mu}, \boldsymbol{\Lambda}, \eta, \nu) = \\
&\quad \frac{|\boldsymbol{\Lambda}|^{-\nu/2}}{2^{\nu P/2}(2\pi)^P \Gamma_P(\nu/2)} \left|\boldsymbol{\Gamma}\right|^{(\nu-P)/2} \cdot \exp\left(-\tfrac{1}{2}\left[\operatorname{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}) + \eta(\boldsymbol{\mu}-\boldsymbol{m})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{\mu}-\boldsymbol{m})\right]\right). \quad \text{(B.4)}
\end{aligned}$$

B.3.2 Loi *a posteriori* conditionnelle des paramètres de classe

Soient données N observations $\boldsymbol{x}_n$, $n = 1, \ldots, N$.[3] Chaque observation est un vecteur de dimension P qui suit une loi multinormale $\mathcal{N}_P(\boldsymbol{x}_n; \boldsymbol{m}, \boldsymbol{\Gamma})$ de moyenne $\boldsymbol{m}$ et de précision $\boldsymbol{\Gamma}$. La loi pour l'ensemble des observations $\boldsymbol{x}_{1:P}$ est le produit des P lois individuelles, et constitue la vraisemblance associée au couple $(\boldsymbol{m}, \boldsymbol{\Gamma})$:

$$\mathrm{p}(\boldsymbol{x}_{1:P} \mid \boldsymbol{m}, \boldsymbol{\Gamma}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{m}, \boldsymbol{\Gamma}) = (2\pi)^{-N/2} \left|\boldsymbol{\Gamma}\right|^{N/2} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{m})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{x}_n - \boldsymbol{m})\right). \quad \text{(B.5)}$$

3. Le terme « observation » n'est pas correct au vu de la hiérarchie, on devrait parler d'observation conditionnelle ou observation probabiliste puisque la variable qui sert d'observation n'est connue qu'indirectement par une inférence probabiliste.

Pour calculer la loi *a posteriori* conditionnelle du couple $(\boldsymbol{m}, \boldsymbol{\Gamma})$, on écrit le produit de la loi *a priori* avec la fonction de vraisemblance :

$$
\begin{aligned}
\mathrm{p}(\boldsymbol{m}, \boldsymbol{\Gamma} \mid \boldsymbol{x}_{1:N}) &\propto \mathrm{p}(\boldsymbol{m}, \boldsymbol{\Gamma})\, \mathrm{p}(\boldsymbol{x}_{1:N} \mid \boldsymbol{m}, \boldsymbol{\Gamma}) \\
&\propto |\boldsymbol{\Gamma}|^{(\nu-P)/2} \cdot |\boldsymbol{\Gamma}|^{N/2} \cdot \\
&\qquad \exp\left(-\frac{1}{2}\left[\operatorname{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}) + \eta(\boldsymbol{m}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{m}-\boldsymbol{\mu}) + \sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{m})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{x}_n-\boldsymbol{m})\right]\right) \\
&= |\boldsymbol{\Gamma}|^{(\nu-P+N)/2} \exp\left(-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma})\right) \\
&\qquad \exp\left(-\frac{1}{2}\left[\eta(\boldsymbol{m}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{m}-\boldsymbol{\mu}) + \sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{m})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{x}_n-\boldsymbol{m})\right]\right).
\end{aligned}
$$

On peut ici retrouver la forme d'une loi normale-wishartienne. Nous montrerons en effet que la loi normale-wishartienne est conjuguée par une fonction de vraisemblance normale. Pour cela, identifions les paramètres *a posteriori*.

Degrés de liberté Nous voyons immédiatement que le nombre de degrés de liberté *a posteriori* vaut

$$\nu^{\mathrm{post}} = \nu + N. \tag{B.6a}$$

Moyenne et nombre d'échantillons

En considérant le dernier terme du produit, nous pouvons identifier la forme d'une loi normale de moyenne $\boldsymbol{\mu}^{\mathrm{post}}$ et de précision $(\eta + N)\boldsymbol{\Gamma} = \eta^{\mathrm{post}}\boldsymbol{\Gamma}$. En utilisant une des méthodes d'identification exposées ci-dessus, on trouve

$$\boldsymbol{\mu}^{\mathrm{post}} = \frac{\eta\boldsymbol{\mu} + N\bar{\boldsymbol{x}}}{\eta + N} \quad \text{avec } \bar{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n \text{ et} \tag{B.7}$$

$$\eta^{\mathrm{post}} = \eta + N. \tag{B.8}$$

Remarque (Résidu) *Nous avons trouvé que* $\eta(\boldsymbol{m}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{m}-\boldsymbol{\mu}) + \sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{m})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{x}_n-\boldsymbol{m})$ *est additivement proportionnel à* $Q(\boldsymbol{m}) = \eta^{\mathrm{post}}(\boldsymbol{m}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{m}-\boldsymbol{\mu})$ *Le terme de proportionalité additif est noté* $X(\boldsymbol{\Gamma})$ *et ne dépend pas de* $\boldsymbol{m}$ *puisque l'influence de ce paramètre est entièrement reproduite dans* $Q(\boldsymbol{m})$*. Pour le calcul de* $X(\boldsymbol{\Gamma})$*, il suffit de soustraire les potentiels et on obtient*

$$X(\boldsymbol{\Gamma}) = \frac{N\eta}{N+\eta}(\boldsymbol{\mu}-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{\mu}-\bar{\boldsymbol{x}}) + \sum_{n=1}^{N}(\boldsymbol{x}_n-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{x}_n-\bar{\boldsymbol{x}}).$$

Nous aurons besoin de cette expression dans le calcul de la matrice d'échelle a posteriori.

Matrice d'échelle

Avec les résultats précédents, nous pouvons exprimer la loi *a posteriori* conditionnelle pour $(\boldsymbol{m}, \boldsymbol{\Gamma})$ de la manière suivante :

$$\mathrm{p}(\boldsymbol{m}, \boldsymbol{\Gamma}) \propto |\boldsymbol{\Gamma}|^{(\nu^{\mathrm{post}}+P)/2} \exp\left(-\tfrac{\eta^{\mathrm{post}}}{2}(\boldsymbol{m}-\boldsymbol{\mu}^{\mathrm{post}})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{m}-\boldsymbol{\mu}^{\mathrm{post}})\right) \cdot \exp\left(-\tfrac{1}{2}(\operatorname{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}) + X(\boldsymbol{\Gamma})\right).$$

Développons le potentiel de la dernière exponentielle :

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}) + X(\boldsymbol{\Gamma}) &= \operatorname{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}) + \tfrac{N\eta}{\eta^{\text{post}}} \underbrace{(\boldsymbol{\mu}-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{\mu}-\bar{\boldsymbol{x}})}_{\in\mathbb{R}} + \sum_{n=1}^{N} \underbrace{(\boldsymbol{x}_n-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{x}_n-\bar{\boldsymbol{x}})}_{\in\mathbb{R}} \\
&= \operatorname{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}) + \tfrac{N\eta}{\eta^{\text{post}}} \operatorname{tr}\Big((\boldsymbol{\mu}-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{\mu}-\bar{\boldsymbol{x}})\Big) + \sum_{n=1}^{N} \operatorname{tr}\Big((\boldsymbol{x}_n-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{x}_n-\bar{\boldsymbol{x}})\Big) \\
&= \operatorname{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}) + \tfrac{N\eta}{\eta^{\text{post}}} \operatorname{tr}\Big((\boldsymbol{\mu}-\bar{\boldsymbol{x}})(\boldsymbol{\mu}-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\Gamma}\Big) + \sum_{n=1}^{N} \operatorname{tr}\Big((\boldsymbol{x}_n-\bar{\boldsymbol{x}})(\boldsymbol{x}_n-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\Gamma}\Big) \\
&= \operatorname{tr}\left(\left[\boldsymbol{\Lambda}^{-1} + \tfrac{N\eta}{\eta^{\text{post}}}(\boldsymbol{\mu}-\bar{\boldsymbol{x}})(\boldsymbol{\mu}-\bar{\boldsymbol{x}})^{\mathrm{T}} + \sum_{n=1}^{N}(\boldsymbol{x}_n-\bar{\boldsymbol{x}})(\boldsymbol{x}_n-\bar{\boldsymbol{x}})^{\mathrm{T}}\right]\boldsymbol{\Gamma}\right).
\end{aligned}$$

La matrice d'échelle *a posteriori* est donc, par identification avec le terme $\operatorname{tr}((\boldsymbol{\Lambda}^{\text{post}})^{-1}\boldsymbol{\Gamma})$ donnée par

$$\boldsymbol{\Lambda}^{\text{post}} = \left[\boldsymbol{\Lambda}^{-1} + \tfrac{N\eta}{\eta^{\text{post}}}(\boldsymbol{\mu}-\bar{\boldsymbol{x}})(\boldsymbol{\mu}-\bar{\boldsymbol{x}})^{\mathrm{T}} + \sum_{n=1}^{N}(\boldsymbol{x}_n-\bar{\boldsymbol{x}})(\boldsymbol{x}_n-\bar{\boldsymbol{x}})^{\mathrm{T}}\right]^{-1} \tag{B.9}$$

B.3.3 Résumé

Pour le couple de paramètres $(\boldsymbol{m}, \boldsymbol{\Gamma})$, nous avons choisi la loi normale-wishartienne $\mathcal{NW}_P(\boldsymbol{m}, \boldsymbol{\Gamma}; \boldsymbol{\mu}, \boldsymbol{\Lambda}, \eta, \nu)$. Étant donnée la vraisemblance normale

$$\mathrm{p}(\boldsymbol{x}_{1:N} \mid \boldsymbol{m}, \boldsymbol{\Gamma}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{m}, \boldsymbol{\Gamma}),$$

nous avons montré

1. que la loi normale-wishartienne est conjuguée par une vraisemblance multinormale et
2. que ses paramètres *a posteriori* (cf. Éqn. (B.6)) sont
   - $\boldsymbol{\mu}^{\text{post}} = \frac{\eta\boldsymbol{\mu}+N\bar{\boldsymbol{x}}}{\eta+N}$,
   - $\boldsymbol{\Lambda}^{\text{post}} = \left[\boldsymbol{\Lambda}^{-1} + \frac{N\eta}{\eta^{\text{post}}}(\boldsymbol{\mu}-\bar{\boldsymbol{x}})(\boldsymbol{\mu}-\bar{\boldsymbol{x}})^{\mathrm{T}} + \sum_{n=1}^{N}(\boldsymbol{x}_n-\bar{\boldsymbol{x}})(\boldsymbol{x}_n-\bar{\boldsymbol{x}})^{\mathrm{T}}\right]^{-1}$,
   - $\eta^{\text{post}} = \eta + N$,
   - $\nu^{\text{post}} = \nu + N$.

Remarque (Interprétation des paramètres) *Les paramètres a posteriori pour la moyenne et la matrice d'échelle font intervenir des grandeurs empiriques, précisément la moyenne empirique des observations $\bar{\boldsymbol{x}}$ et la matrice de covariance empirique $\sum_{n=1}^{N}(\boldsymbol{x}_n-\bar{\boldsymbol{x}})^{\mathrm{T}}(\boldsymbol{x}_n-\bar{\boldsymbol{x}})$. C'est ainsi que les informations fournies par les observations sont intégrées et ajoutées aux connaissances a priori dans la loi a posteriori.*

Remarque (Échantillonnage) *La loi a posteriori du couple étant séparable comme toute loi normale-wishartienne en une partie* moyenne *et une partie* précision, $\mathrm{p}(\boldsymbol{m}, \boldsymbol{\Gamma} \mid \boldsymbol{x}_{1:N}) = \mathrm{p}(\boldsymbol{\Gamma} \mid \boldsymbol{x}_{1:N}) \cdot \mathrm{p}(\boldsymbol{m} \mid \boldsymbol{\Gamma}, \boldsymbol{x}_{1:N})$, *l'échantillonnage du couple peut se faire séparément, tout en respectant la dépendance des deux paramètres. Pour cela, on tire aléatoirement un échantillon pour la précision $\boldsymbol{\Gamma}$ sous sa loi a posteriori conditionnelle (par rapport aux données). Ensuite, avec cette connaissance, on tire un échantillon aléatoire pour la moyenne $\boldsymbol{m}$ sous sa loi a posteriori conditionnelle (par rapport aux données* **et** *à la précision). Ainsi, nous obtenons une réalisation aléatoire jointe du couple $(\boldsymbol{m}, \boldsymbol{\Gamma})$.*

Remarque (Cas limite) *Considérons un cas limite de la loi normale-wishartienne. Pour cela, il suffit de faire tendre le nombre de degrés de liberté ν vers 0, le déterminant de la matrice d'échelle $|\mathbf{\Lambda}|$ vers ∞, et enfin le nombre d'échantillons a priori η vers 0. Ainsi, la loi normale-wishartienne tend vers une extension multidimensionnelle de la loi de Jeffreys associée à une vraisemblance normale multivariée :*

$$\mathrm{p}(\boldsymbol{m}, \mathbf{\Gamma}) \propto |\mathbf{\Gamma}|^{-P/2}. \tag{B.10}$$

Dans ce cas, il est intéressant d'observer les paramètres a posteriori. En effet, le nombre d'échantillons intervenant a posteriori est naturellement N, et la même chose est vraie pour le degré de liberté. Que se passe-t-il pour les paramètres moyenne $\boldsymbol{\mu}^{\mathrm{post}}$ et matrice d'échelle $\mathbf{\Lambda}^{\mathrm{post}}$? Le nombre d'échantillons a priori tendant vers 0, la moyenne a posteriori de la distribution est simplement la moyenne empirique des observations, $\boldsymbol{\mu}^{\mathrm{post}} = \bar{\boldsymbol{x}}$. Ensuite, avec le même raisonnement et sachant que $\mathbf{\Lambda}^{-1} \to 0$, la matrice d'échelle a posteriori tend vers l'inverse de la covariance empirique, ou vers la « précision empirique » Π. Nous retrouvons donc les « estimateurs classiques » dans le calcul bayésien des paramètres des classes.

Quant aux paramètres de la loi a posteriori du couple des paramètres, l'échantillonnage se fait d'abord pour la matrice de précision $\mathbf{\Gamma}$ suivant la loi wishartienne qui prend comme matrice d'échelle la précision empirique et comme nombre de degrés de liberté $N+1$, la moyenne $\boldsymbol{m}$ suivant une loi normale qui prend comme paramètres la moyenne empirique des observations et la précision apportée par N observations, $N\mathbf{\Gamma}$:

$$\begin{aligned} \mathbf{\Gamma} \mid \boldsymbol{x}_{1:N} &\sim \mathcal{W}_P(\Pi, N+1), \\ \boldsymbol{m} \mid \mathbf{\Gamma}, \boldsymbol{x}_{1:N} &\sim \mathcal{N}_P(\bar{\boldsymbol{x}}, N\mathbf{\Gamma}). \end{aligned}$$

B.4 Loi a posteriori conditionnelles

Dans ce document, un certain nombre de lois *a posteriori* ont dû être calculées. La plupart d'entre elles sont issues d'une conjugaison de la loi *a priori* correspondante par la vraisemblance associée. Il s'agit notamment de lois normales, gamma, ou normales-wishartiennes. Nous avons détaillé les calculs génériques pour elles précédemment dans ce chapitre. En remplaçant la variable dans le calcul par la variable d'intérêt, les résultats sont immédiats.

Pour les paramètres chromatographiques, distribués *a priori* sous des lois uniformes, une telle conjugaison n'est pas possible, comme montre un calcul d'isolation de paramètres (*cf.* l'annexe de [1]). L'expression de ces lois ne peut être simplifiée ou mise dans des formes standard.

C Choix de modèle

C.1 Marginalisation par Moyenne Harmonique

Soit θ un paramètre de nuisance (potentiellement vectoriel) dans un problème de sélection de modèle. On se donne $\varphi(\theta)$ une mesure de probabilité. On note $c \in \mathcal{C}$ représentatif pour la classe/le modèle, et $\mathbf{Y}$ les données.

Dans le cadre du calcul de la loi *a posteriori* $\Pr(c \mid \mathbf{Y})$, nous pouvons restreindre nos efforts au le calcul de l'évidence $p(\mathbf{Y} \mid c)$ puisque dans la règle de Bayes $\Pr(c \mid \mathbf{Y}) = \frac{\Pr(c)\, p(\mathbf{Y}|c)}{p(\mathbf{Y})}$, l'*a priori* de la classe est connue, et le facteur de normalisation $p(\mathbf{Y})$ s'obtient par la somme des probabilités *a posteriori* et peut ainsi être déduite.

La présentation [99] donne pour l'évidence l'identité suivante :

$$\frac{1}{p(\mathbf{Y} \mid c)} = \mathbb{E}_{\theta|\mathbf{Y},c}\left(\frac{\varphi(\theta)}{p(\mathbf{Y}, \theta \mid c)}\right). \tag{C.1}$$

Écrivons cette espérance *a posteriori* de deux manières différentes.

C.1.1 Espérance, calcul continu

Avec les notations précédentes, montrons que l'évidence est indépendante du choix de φ :

$$\begin{aligned}
C_\infty(\varphi) &= \mathbb{E}_{\theta|\mathbf{Y},c}\left(\frac{\varphi(\theta)}{p(\mathbf{Y}, \theta \mid c)}\right) \\
&= \int \frac{\varphi(\theta)}{p(\mathbf{Y}, \theta \mid c)}\, p(\theta \mid \mathbf{Y}, c)\, d\theta \\
&= \int \frac{\varphi(\theta)}{p(\mathbf{Y}, \theta \mid c)} \frac{p(\theta, \mathbf{Y} \mid c)}{p(\mathbf{Y} \mid c)}\, d\theta \\
&= \int \frac{\varphi(\theta)}{p(\mathbf{Y} \mid c)}\, d\theta \\
&= \frac{1}{p(\mathbf{Y} \mid c)} \int \varphi(\theta)\, d\theta \\
&= \frac{1}{p(\mathbf{Y} \mid c)} = C_\infty.
\end{aligned} \tag{C.2}$$

Ainsi, quel que soit la mesure φ, l'évidence est de valeur C_∞. Autrement dit, l'évidence est invariante aux changements de mesure de probabilité.

C.1.2 Espérance, calcul discret

Soit $\theta^{(k)} \sim \mathrm{p}(\theta \mid \mathbf{Y}, \mathrm{c})$ pour $k = 1, \ldots, K$ des échantillons de la loi *a posteriori* du paramètre θ conditionnée par la classe c. Alors,

$$\begin{aligned} C_\infty(\varphi) &= \mathbb{E}_{\theta|\mathbf{Y},\mathrm{c}}\left(\frac{\varphi(\theta)}{\mathrm{p}(\mathbf{Y}, \theta \mid \mathrm{c})}\right) \\ &= \lim_{K\to\infty}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{\varphi(\theta^{(k)})}{\mathrm{p}(\mathbf{Y} \mid \theta^{(k)})\,\mathrm{p}(\theta^{(k)} \mid \mathrm{c})}\right] \\ &= \lim_{K\to\infty}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{\varphi\left((\boldsymbol{\theta}^{\mathrm{bio}})^{(k)}, (\boldsymbol{\theta}^{\mathrm{inst}})^{(k)}\right)}{\mathrm{p}(\mathbf{Y} \mid (\boldsymbol{\theta}^{\mathrm{inst}})^{(k)})\,\mathrm{p}\left((\boldsymbol{\theta}^{\mathrm{inst}})^{(k)} \middle| (\boldsymbol{\theta}^{\mathrm{bio}})^{(k)}\right)\mathrm{p}\left((\boldsymbol{\theta}^{\mathrm{bio}})^{(k)} \middle| \mathrm{c}\right)}\right] = \lim_{K\to\infty} C_K(\varphi) \end{aligned} \tag{C.3}$$

où la dernière ligne tient à cause de la structure hiérarchique des paramètres.

Le choix de la fonction $\varphi(\theta)$ détermine la vitesse de convergence de la suite $C_K(\varphi)$.

C.1.3 Mise en commun des résultats

Au vu des deux expressions établies, nous avons

$$C_\infty = C_\infty(\varphi) = \lim_{K\to\infty} C_K(\varphi) \tag{C.4}$$

l'équation sur laquelle est basée le calcul numérique de l'évidence. Il est fait à partir des échantillons *a posteriori*, obtenus par exemple par un algorithme de type MCMC.

Il ne reste plus que le choix de $\varphi(\theta)$. En effet, on peut choisir pour cette densité la loi uniforme sur l'espace de probabilité de θ ce qui est probablement le choix le plus simple. Cependant, dans un scénario où l'espace n'est pas borné, l'intégrale sur la fonction φ n'est plus fini, et le passage de l'avant-dernière à la dernière ligne dans Éqn. (C.2) n'est plus correct.

Un autre choix raisonnable pour φ est l'utilisation de la loi *a priori* de θ, $\mathrm{p}(\theta \mid \mathrm{c})$. On a alors

$$C_K(\varphi) = C_K(\mathrm{p}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{\mathrm{p}\left(\mathbf{Y}|\theta^{(k)}\right)}, \tag{C.5}$$

les échantillons étant toujours tirés sous la loi *a posteriori*. Dit autrement, avec $\varphi(\theta) = \mathrm{p}(\theta)$, l'inverse de l'évidence se calcule comme moyenne des inverses des vraisemblances totales. Ceci correspond au choix que nous avons fait pour l'établissement de l'Éqn. (5.18) (page 113).

Remarque *Si, au lieu de prendre l'espérance sous la loi a posteriori de θ dans la définition Éqn. (C.1), on calculait l'espérance sous la loi a priori du paramètre, on arriverait à un résultat similaire, à savoir l'approximation de l'évidence par la moyenne arithmétique des vraisemblances :*

$$\mathrm{p}(\mathbf{Y} \mid \mathrm{c}) = \mathbb{E}_{\theta|\mathrm{c}}\left(\frac{\varphi(\theta)}{\mathrm{p}(\mathbf{Y}, \theta \mid \mathrm{c})}\right) = \lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}\mathrm{p}(\mathbf{Y} \mid \theta, \mathrm{c}).$$

Numériquement, cela voudrait dire que l'on tire des échantillons sous l'a priori de θ qui peut être très large, la plupart des valeurs ayant un apport quasi nul au résultat ; la zone à forte probabilité a posteriori qui y apporte le plus est possiblement sous-échantillonnée, certaines valeurs ne sont pas prise en compte, à moins de faire un échantillonnage ultra-fin nécessitant une puissance calculatoire énorme.

C.2 Reversible Jump MCMC

Dans le cas du modèle binaire dans la Rq. ⟨5.9⟩, explicitons les probabilités d'acception dans le cas où le dernier modèle choisi est $\mathrm{c}^{(k-1)} = \mathrm{S}$ et $\mathrm{c}^{(k-1)} = \mathrm{M}$. On note pour cela

$$\mathcal{M}_\mathrm{S} : \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}_\mathrm{S}, \boldsymbol{\Gamma}_\mathrm{S}) \tag{$\mathcal{M}_\mathrm{S}$}$$
$$\mathcal{M}_\mathrm{M} : \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}_\mathrm{M}, \boldsymbol{\Gamma}_\mathrm{M}) \tag{$\mathcal{M}_\mathrm{M}$}$$

Par soucis de simplicité, considérons uniquement le cas scalaire dans lequel nous définissons

$$\mathcal{M}'_\mathrm{S} : x \sim \mathcal{N}(m_\mathrm{S}, \gamma_\mathrm{S}) \tag{$\mathcal{M}'_\mathrm{S}$}$$
$$\mathcal{M}'_\mathrm{M} : x \sim \mathcal{N}(m_\mathrm{M}, \gamma_\mathrm{M}). \tag{$\mathcal{M}'_\mathrm{M}$}$$

L'algorithme *Reversible Jump Monte Carlo Markov Chain* (RJMCMC) étend l'algorithme MCMC au choix de modèle [21, Sect. 11.2]. L'étape du choix de modèle fonctionne comme une étape de Metropolis-Hastings. À chaque itération, l'algorithme met en concurrence deux modèles : le modèle choisi à la dernière itération, et un modèle proposé, tiré sous la loi *a priori* des modèles. Ensuite, pour décider du modèle de l'itération suivante, une probabilité d'acceptation est évaluée dans laquelle se trouve les probabilités *a priori* et les vraisemblances du dernier échantillon joint par rapport aux modèles. Seulement, les modèles ne sont pas toujours de la même dimension ou de même support. Il convient alors de définir des fonctions de transformation $T(\cdot)$ du premier au second modèle en question, qui trouve également son influence dans la probabilité d'acceptation. Dans le cas de même dimension, cette probabilité s'écrit

$$\delta_{\mathrm{S}\to\mathrm{M}} = \min\left(1, \frac{\mathrm{p}(\mathrm{M}, x_\mathrm{M})}{\mathrm{p}(\mathrm{S}, x_\mathrm{S})} \frac{\pi_{\mathrm{M}\to\mathrm{S}}}{\pi_{\mathrm{S}\to\mathrm{M}}} \left|\frac{\partial T(x_\mathrm{S})}{\partial x_\mathrm{S}}\right|\right),$$

où $\pi_{\mathrm{c}_i\to\mathrm{c}_j}$ est la probabilité pour choisir un saut du modèle c_i au modèle c_j et $\mathrm{p}(\mathrm{c}, x_\mathrm{c})$ le produit des lois *a priori* $\Pr(\mathrm{c}) \cdot \mathrm{p}(x_\mathrm{c} \mid \mathrm{c})$ [21, Sect. 11.1.1]. Dans [91, Sect. 6.7.2] cependant (certainement pour des raisons d'efficacité que nous avons mentionnées pour le calcul de l'évidence, Rq. 1), les auteurs proposent de substituer $\mathrm{p}(\mathrm{c}, x_\mathrm{c})$ par la loi *a posteriori* $\mathrm{p}(\mathrm{c}, x_\mathrm{c} \mid \mathbf{Y}) \propto \Pr(\mathrm{c})\, \mathrm{p}(x_\mathrm{c} \mid \mathrm{c}) \ell_\mathrm{c}(x_\mathrm{c})$ dépendant de l'état, tout du moins dans le cas de mélanges gaussiens. ℓ_c désigne la vraisemblance pour le modèle c.

Considérons alors un saut du modèle S au modèle M, et définissons la transformation nécessaire. Elle doit vérifier la bijectivité entre les modèles et doit également fusionner ou séparer les informations du modèle courant d'une manière « raisonnable » pour l'utilisation dans le modèle proposé. Dans le cas monodimensionnel que nous étudierons ici, nous choisissons une transformation affine :

$$(x_\mathrm{M}) = T(x_\mathrm{S}) = (x_\mathrm{S} + \Delta m)$$

avec Δm connu. La jacobienne, ici la dérivée devient

$$\frac{\partial}{\partial(x_\mathrm{S})} T(x_\mathrm{S}) = 1$$

dont le déterminant vaut 1. Alors, la probabilité d'acceptation devient

$$\delta = \min\left(1, \frac{\mathrm{p}(\mathrm{M}, x_\mathrm{M})}{\mathrm{p}(\mathrm{S}, x_\mathrm{S})} \frac{\pi_{\mathrm{M}\to\mathrm{S}}}{\pi_{\mathrm{S}\to\mathrm{M}}}\right).$$

En admettant une probabilité de saut uniforme, *i.e.* $\pi_{\mathrm{S}\to\mathrm{M}} = \pi_{\mathrm{M}\to\mathrm{S}}$, la probabilité d'acceptation devient le rapport des probabilités $\mathrm{p}(\mathrm{M}, x_{\mathrm{M}})/\mathrm{p}(\mathrm{S}, x_{\mathrm{S}})$.

Allons encore un peu plus loin et proposons $\Delta m = 0$, *i.e.* quel que soit le modèle de départ et celui d'arrivée, la transformation de la concentration protéique est l'identité. Ainsi, la probabilité d'acception est le rapport $\mathrm{p}(\mathrm{M}, \boldsymbol{x})/\mathrm{p}(\mathrm{S}, \boldsymbol{x}) = \mathrm{Pr}(\mathrm{M})/\mathrm{Pr}(\mathrm{S}) \cdot \mathrm{p}(\boldsymbol{x} \mid \mathrm{M})/\mathrm{p}(\boldsymbol{x} \mid \mathrm{S})$. Finalement, en utilisant la loi *a posteriori* $\mathrm{p}(\mathrm{c}, x_{\mathrm{c}} \mid \mathbf{Y}) \propto \mathrm{Pr}(\mathrm{c})\, \mathrm{p}(x) \ell_{\mathrm{c}}(x)$ au lieu de l'*a priori* dans l'expression (en suivant l'idée donnée précédemment), l'approche RJMCMC avec les restrictions et limites introduites coïncide avec celle présentée dans la Sect. 5.4.3.

D Quantification LC-Full-MS

Nous donnons ici les résultats de simulation obtenus pour l'utilisation du couplage LC-Full-MS, la section Sect. 4.5 se concentrant sur le mode SRM dû à un plus grand nombre de données.

Pour tester notre méthode d'inversion bayésienne hiérarchique (IBH), nous mettons en place des échantillons biologiques simulés. Nous la comparerons à l'inversion bayésienne (non hiérarchique, notée IB [1]) et à une version modifiée du maximum du pic. Cette dernière méthode que nous noterons MP+ utilise l'estimation de la position du pic chromatographique issue des résultats de l'IBH, dans le voisinage de laquelle la valeur maximale est choisie.

Avant de décrire plus précisément les données, nous introduirons les critères d'évaluation qui permettent de valider notre méthode d'inversion et de comparer les méthodes.

Nous utilisons comme critères de performance la droite de régression, l'erreur quadratique moyenne par rapport à la valeur vraie ($\mathbb{EQ}_{\boldsymbol{x},\mathbf{Y}}(\psi^{\mathrm{q}})$ au sens de Sect. 2.8) et l'écart quadratique moyen entre estimation et prédiction par la droite de régression,

$$\mathbb{J}_{\boldsymbol{x},\mathbf{Y},\psi^{\mathrm{q}}}(\alpha,\beta) = \mathbb{E}_{\boldsymbol{x},\mathbf{Y}}\left((\psi^{\mathrm{q}}(\mathbf{Y}) - \alpha - \beta\boldsymbol{x})^2\right).$$

Cette expression correspond aussi au critère à minimiser dans la recherche des coefficient de régression.

D.1 Description des données

Pour démontrer l'apport d'une structure hiérarchique [119], nous avons enrichi notre modèle pour comprendre un processus de digestion aléatoire, comme nous l'avons présenté dans le Ch. 3. Pour cela, nous ciblons une protéine, Neuron Specific Enolase (NSE), qui a plusieurs peptides. Nous la supposons bipeptidique avec les peptides ① et ②, donnant lieu à un polypeptide ①②. Les signatures élémentaires des peptides sont définies par les proportions de charges et proportions de neutrons, ainsi que les profils spectrométriques et chromatographiques dont les paramètres (hormis le temps de rétention) ont été appris par calibrage externe.

Nous simulons quatre cohortes test qui comportent 60 échantillons chacune. Chaque cohorte est caractérisée par la moyenne m autour de laquelle la concentration protéique est distribuée. Les valeurs pour m sont 50, 100, 200 et 400 (unité arbitraire). Pour toutes les cohortes, la concentration de la protéine marquée est distribuée autour de 150. La précision de ces distributions vaut $\gamma = (0.1m)^{-1}$.

Supposons une préparation idéale, *i.e.* $\psi = 1$. La digestion est décrite par la matrice de digestion $\mathbf{D} = \begin{bmatrix}1 & 1 & 1\end{bmatrix}^{\mathrm{T}}$. Ensuite, pour simuler une digestion incomplète, $\chi = \chi_1 = \chi_2$ est tiré aléatoirement et uniformément dans l'ensemble $\{0.4, 0.55, 0.7, 0.85, 1\}$.

Les peptides sont séparés suivant leurs temps de rétention $\boldsymbol{\tau}$. La loi de simulation est uniforme : $\boldsymbol{\tau} \sim \mathcal{U}([16.2, 16.8] \times [21.2, 22.6] \times [16.0, 16.4])$.

Le paramètre de gain de système par peptides, $\boldsymbol{\zeta}$, est distribué sous une loi normale de moyenne $[10, 20, 25]$ et de précision $\mathrm{diag}(10, 10, 10)$.

Tab. D.1 Paramètres et distributions utilisés pour la simulation des données.

- protéine
  - une protéine bipeptidique ①②
  - protéine native
    - $x \sim \mathcal{N}(m, \gamma)$ où $m \in \{50, 100, 200, 400\}$ et $\gamma = 1/(0.1m)$
  - concentration de la protéine marquée 50
- peptide
  - deux monopeptides ① et ②, un polypeptide ①②
  - précision de digestion $\mathbf{\Gamma}_{\boldsymbol{\kappa}} = \mathrm{diag}(100, 100)$
  - rendement de digestion
    - $\chi \sim \mathcal{U}(\{0.4, 0.55, 0.7, 0.85, 1\})$
  - masses
    - $m_1 = 580.77\,\mathrm{Da}$, $m_1^* = m_1 + 9\,\mathrm{Da}$
    - $m_2 = 995.47\,\mathrm{Da}$, $m_2^* = m_2 + 9\,\mathrm{Da}$
    - $m_{12} = m_1 + m_2$, $m_{12}^* = m_{12} + 9\,\mathrm{Da}$
  - charges $J = 2$
    - $\pi_{i,1} = 0, i \in \{1, 2, 12\}$
    - $\pi_{i,2} = 1, i \in \{1, 2, 12\}$
  - neutrons $K = 0$
    - $\pi_{i,1,0} = 0, i \in \{1, 2, 12\}$
    - $\pi_{i,2,0} = 1, i \in \{1, 2, 12\}$
  - temps de rétention
    - $\tau_1 \in [16.2, 16.8]$ min
    - $\tau_2 \in [21.2, 22.6]$ min
    - $\tau_{12} \in [16.0, 16.4]$ min
  - gain d'ionisation
    - $\xi_1 \sim \mathcal{N}(10, 10)$
    - $\xi_2 \sim \mathcal{N}(20, 10)$
    - $\xi_{12} \sim \mathcal{N}(25, 10)$
- observations
  - largeur des pics
    - spectrométriques : $\lambda = (0.08)^{-1}$
    - chromatographiques : $\lambda = 6^{-1}$
  - fréquence d'échantillonnage
    - spectrométrique : $f^{\mathrm{e}}_{\mathscr{S}} = 1/M^{\mathrm{e}}_{\mathscr{S}} = 1/0.04$
    - chromatographique : $f^{\mathrm{e}}_{\mathscr{C}} = 1/T^{\mathrm{e}}_{\mathscr{C}} = 1/3$
  - bruit de mesure
    - bruit blanc gaussien de moyenne nulle et de précision γ
    - $\gamma \sim \mathcal{U}(\{0.004, 0.008, 0.016\})$

Finalement, l'inverse-variance du bruit de mesure est choisie uniformément dans l'ensemble $\{0.004, 0.008, 0.016\}$. Nous avons volontairement simulé un bruit relativement fort pour nous mettre dans un cadre réaliste et pour montrer l'apport de la méthodologie bayésienne par rapport au MP+.

Toutes les valeurs nécessaires et utilisées pour la simulation des données sont résumées dans la Tab. D.1.

D.2 Résultats et comparaison

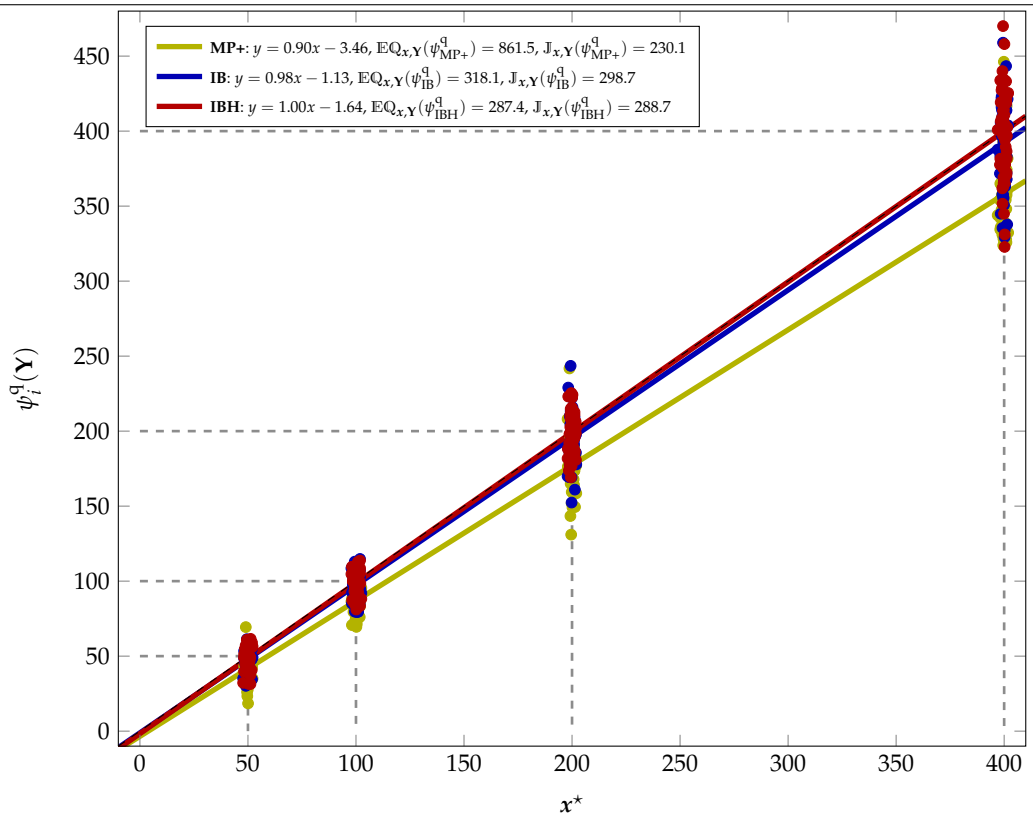
Les choix algorithmiques pour les méthodes IBH et IB ont été les suivants. Les lois *a priori* des paramètres $\boldsymbol{x}$, $\boldsymbol{\xi}$ et γ étaient faiblement informatives. Nous ne monitorons pas la convergence par le lancement de plusieurs chaînes MCMC par échantillons biologiques ; pour quand même être relativement sûr de converger, nous optons pour un nombre d'itérations élevé, à savoir $K = 5000$, avec un temps de chauffe de $K_0 = 4000$. Ainsi, les estimations seront faites à partir des derniers 1000 échantillons des lois.

Comparons les résultats obtenus avec les trois méthodes mentionnées. Pour cela, considérons les droites de régression dans la Fig. D.1 où nous avons tracé sur l'abscisse la valeur vraie et sur l'ordonnée la valeur estimée associée. Les points jaunes, bleus et rouges sont les estimations du MP+, de l'IB et de l'IBH respectivement, idem pour les droites de régression.

Nous pouvons voir que les trois méthodes ont des performances similaires : plus la concentration vraie est élevée, plus la dispersion est grande, mais les estimations sont centrées autour d'un point moyen qui est proche du point $(\boldsymbol{x}^\star, \boldsymbol{x}^\star)$ sur la bissectrice. La dispersion par rapport à la droite de régression est donnée par $\mathbb{J}_{\boldsymbol{x},\mathbf{Y}}(\psi_i^{\mathrm{q}})$ qui vaut 230.1, 298.7, 288.7 pour $i \in \{\text{MP+, IB, IBH}\}$ respectivement. Nous trouvons donc la variation des estimations la moins importante par rapport à la droite de régression chez le MP+.

L'erreur quadratique moyenne, $\mathbb{EQ}_{\boldsymbol{x},\mathbf{Y}}(\psi_i^{\mathrm{q}})$, est de 287.4 pour l'IBH, légèrement plus élevée pour l'IB avec 318.1, et de 861.5 pour le MP+ du fait que cette dernière méthode livre la moins bonne estimation des trois, l'IBH fournissant les meilleurs résultats. Si on

Fig. D.1 Droites de régression pour les estimations par les méthodes IBH, IB et MP+.



suppose un biais moyen nul (ce qui est correct pour IB et IBH grâce à la quantification selon l'EAP, mais pas forcément correct pour MP+), l'écart moyen de l'estimation par rapport à la valeur vraie est d'environ 17.0, 17.3 et 29.3 pour les trois méthodes respectivement. Les coefficients de variation valent respectivement 4.25%, 4.325% et 7.325% ce qui reste remarquable du point de vue d'un biologiste. ↓1

Quant aux pentes des droites de régression, elles valent 1 pour l'IBH, 0.98 pour l'IB et 0.9 pour MP+. Les estimations dans leur ensemble sont bonnes, avec un avantage pour l'IBH, suivi par l'IB. Rappelons que nous n'avons pas introduit de concentration endogène dans les échantillons biologiques simulés. Au vu des droites de régression, la valeur au point $x^{\star} = 0$ est proche de 0 pour chacune des méthodes.

D.3 Discussion

Les méthodes montrent de bons résultats. On peut s'étonner des performances du MP+ malgré l'erreur quadratique moyenne très supérieure par rapport à IBH.. Il est censé être très sensible au bruit qui a été choisi très fort ici dans nos simulations. N'oublions néanmoins pas que la position du pic chromatographiques de référence pour les calculs a été estimée avec l'inférence bayésienne que nous avons présentée. De plus, l'intensité n'a pas été estimée manuellement, mais déterminée directement par la valeur de la donnée à la position spécifiée.

La méthode d'inversion bayésienne hiérarchique prend en compte l'information que fournit le polypeptide, donc de l'information issue d'une mauvaise digestion. Notons

1. Une pente de la droite de régression entre 0.9 et 1.1 ainsi qu'un coefficient de variation de 10% sont normalement admissibles pour la validation d'une méthode de quantification en protéomique.

que cette information est disponible dans les deux cas : la donnée comporte le peptide mal digéré, et l'information sur sa masse et son temps de rétention moyen peut être calculée. Au vu des meilleurs résultats par rapport à l'inversion bayésienne non hiérarchique, nous pouvons conclure que l'ajout de l'information apporte une amélioration de l'estimation. Ceci n'est pas un fait nouveau, mais connu (plus on a de données, meilleure est l'estimation). Elle nous conforte néanmoins dans la pensée que la structure hiérarchique des données apporte un gain et de la robustesse puisque les processus peuvent être modélisés avec plus de précision, d'exactitude, de rigueur et de facilité.

E Notes biographiques

[i] Thomas Bayes, né en 1702 à Londres, Angleterre, mort le 17 avril 1761 à Tunbridge Wells, Kent, Angleterre. Mathématicien britannique et pasteur de l'Église presbytérienne. Fondateur de « probabilité inverse », méthodologie de statistiques qui est connue aujourd'hui sous le nom d'*inférence bayésienne*.

[ii] Pierre-Simon Marquis de Laplace, né le 23 mars 1749 à Beaumont-en-Auge, Normandie, France, mort le 5 mars 1827 à Paris, France. Mathématicien et astronome français, personnage important de la théorie de la probabilité, ayant accepté et généralisé les propositions de Th. Bayes.

[iii] Sir Harold Jeffreys, né le 22 avril 1891 à Fatfield, Durham, Angleterre, mort le 18 mars 1989 à Cambridge, Angleterre. Mathématicien, statisticien, astronome, (géo-)physicien anglais. Son ouvrage « Theory of Probability » [19] est un des plus cités dans la communauté bayésienne. Il fut fait Chevalier en 1953 par la Reine Elizabeth II.

[iv] Sir Ronald Aylmer Fisher, né le 17 février 1890 à East Finchley, Londres, Angleterre, mort le 29 juillet 1962 à Adélaïde, Australie. Biologiste, eugéniste, évolutionniste et statisticien britanique. Connu notamment en statistique pour la proposition de l'estimation du maximum de vraisemblance (1912/1922), de l'information de Fisher et de l'Analyse de la Variance (ANOVA, 1924). En hommage à ses travaux, il fut fait Chevalier en 1953 par la Reine Elizabeth II en même temps que Sir Harold Jeffreys, même si le décret de nomination de Fisher fut signé en 1952 déjà.

[v] Karl Pearson, né le 27 mars 1857 à Islington, Londres, Angleterre, mort le 27 avril 1936 à Coldharbour, Surrey, Angleterre. Entre autre mathématicien et statisticien britannique, fondateur du premier département de statistiques à l'*University College* à Londres et co-fondateur du journal *Biometrika*, contribua grandement à l'établissement des statistiques mathématiques. Il refusa l'attribution du titre de l'Ordre de l'Empire Britannique ainsi que celui de Chevalier.

[vi] Solomon Kullback, né le 3 avril 1907 à Brooklyn, New York, États-Unis d'Amérique, mort le 05 août 1994 à Boynton Beach, Floride, États-Unis d'Amérique. Cryptanalyste, mathématicien, théoricien de l'information américain. La collaboration avec R. Leibler mena une nouvelle méthode de mesure de ressemblance entre populations, appelée aujourd'hui en leur honneur *divergence de Kullback-Leibler*.

[vii] Richard Leibler, né le 18 mars 1914 à Chicago, Illinois, États-Unis d'Amérique, mort le 25 octobre 2003 à Reston, Virginie, États-Unis d'Amérique. Mathématicien et cryptologue américain. La collaboration avec S. Kullback mena une nouvelle méthode de mesure de ressemblance entre populations, appelée aujourd'hui en leur honneur *divergence de Kullback-Leibler*.

[viii] Andreï Andreïevitch Markov (Андрей Андреевич Марков), né le 2 juin 1856 à Riazan, Empire russe, mort le 20 juillet 1922 à Pétrograd, RSFS de Russie (aujourd'hui Saint-Pétersbourg). Mathématicien russe. Par l'analyse de la succession de lettres d'un ouvrage russe, Markov nota que les lettres suivent certaines contraintes de dépendances par rapport aux lettres précédentes.

[ix] Nicholas Metropolis, né le 11 juin 1915 à Chicago, États-Unis d'Amérique, mort le 17 octobre 1984 à Los Alamos, New Mexico, États-Unis d'Amérique. Physicien gréco-américain. A mené l'équipe de recherche (incluant S. Ulam et J. von Neumann) qui a développé la méthode d'intégration qui est appelée aujourd'hui « méthode d'intégration de Monte-Carlo » (en honneur à l'amour de ses collègues pour le jeu du hasard).

[x] Stanisław Marcin Ulam, né le 13 avril 1909 à Lwów, Pologne (aujourd'hui Lviv, Ukraine), mort le 13 mai 1984 à Santa Fe, New Mexico, États-Unis d'Amérique. Mathématicien américain d'origine polonaise et co-développeur de la bombe à hydrogène, participant au projet Manhattan. Ayant été confronté à un problème d'intégration difficile, il apporta l'idée de la méthode de Monte-Carlo à ses développeurs.

[xi] John von Neumann, né le 28 décembre 1903 à Budapest, Autriche-Hongrie, mort le 8 février 1957 à Washington, D.C., États-Unis d'Amérique. Mathématicien et physicien américano-hongrois. Parmi ses travaux polyvalents, on cite le développement de la méthode de Monte-Carlo. Il a également donné son nom à l'architecture de von Neumann pour les ordinateurs.

[xii] Abou Jafar Mouhammad ben Musa al Khwarizmi (أبو عبد الله محمّد بن موسى الخوارزمي), né vers 780 à Khiva dans la région Khwarezm (aujourd'hui appartenant à l'Ouzbékistan), mort vers 850 à Bagdad. Mathématicien, astronome, géographe perse, membre de la *Maison de la Sagesse* de Bagdad. Son ouvrage *Livre de l'addition et de la soustraction d'après le calcul indien* (كتاب الجامع والتقريق بحساب الهند) fut traduit en langue latine sous le titre *Dixit Algorismi*, d'où l'origine du mot *algorithme*.

[xiii] W. Keith Hastings, né le 21 juillet 1930 à Toronto, Ontaria, Canada, mathématicien canadien. Généralisa les travaux de Metropolis et collaborateurs sur l'algorithme d'échantillonnage de densités de probabilité.

[xiv] Josiah Willard Gibbs, né le 11 février 1839 et mort le 28 avril 1903 à New Haven, Connecticut, États-Unis d'Amérique. Physicien, chimiste, mathématicien américain. Les frères Geman ont appelé l'algorithme d'échantillonnage en son honneur en référence à l'analogie entre l'algorithme et les physiques statistiques.

[xv] Mikhaïl Semionovitch Tsvet (Михаил Семенович Цвет) né le 14 mai 1872 à Asti, Italie, mort ne 26 juin 1919 à Voronej, Russie. Botaniste russe et inventeur de la chromatographie d'adsorption. L'origine du nom « chromatographie » est probablement due au nom de l'inventeur, le mot grec χρωμα (khrôma) désignant la même chose que le mot

russe цвет (cvet) : *couleur*.

[xvi] Paul Dirac, né le 08 août 1902 à Bristol, Angleterre, mort le 20 octobre 1984 à Tallahassee, Floride, États-Unis d'Amérique. Physicien théorique britannique, prix Nobel de physique en 1933 (avec E. Schrödinger) pour les avancées dans la théorie atomique et co-fondateur de la physique quantique.

[xvii] Ferdinand Georg Frobenius, né le 26 octobre 1849 à Berlin, Allemagne, mort le 03 août 1917 à Charlottenburg, Berlin, Allemagne. Mathématicien allemand, ses travaux reposent principalement sur la théorie des groupes et la représentation de groupe. Beaucoup de notions mathématiques portent son nom, comme l'homomorphisme, le groupe, la matrice et la norme de Frobenius. Cette dernière est aussi appelé norme de Schur, en hommage au mathématicien Issai Schur qui était un étudiant de Frobenius.

[xviii] John Wishart, né le 28 novembre 1898 à Montrose, Écosse, mort le 14 juillet 1956 à Acapulco, Méxique. Statisticien écossais et mathématicien, assistant de Pearson et Fisher, proposa en 1928 la distribution qui porte aujourd'hui son nom.

Liste des figures, tables et algorithmes

Liste des figures

Liste des tables

Liste des algorithmes

Références

[1] G. Strubel, « Reconstruction de profils moléculaires : modélisation et inversion d'une chaîne de mesure protéomique, » thèse de doctorat, École Polytechnique de Grenoble, 2008.

[2] C. P. Robert, *The Bayesian Choice : From Decision-Theoretic Foundations to Computational Implementation*, 2ème édition. Springer-Verlag New York Inc., 2007.

[3] A. Gelman, J. B. Carlin, H. S. Stern, et D. B. Rubin, *Bayesian Data Analysis*, 2ème édition, dans la série *Texts in Statistical Science*. Chapman & Hall/CRC, juillet 2003.

[4] J. M. Berg, J. L. Tymoczko, et L. Stryer, *Biochemistry*, 7ème édition. W. H. Freeman, décembre 2010.

[5] B. Jordan, M. M. Perez-Perez, R. Ossig, S. Kaforou, E. Mery, L. Brunet-Errard, C. Paulus, M. Kritsotakis, L. Gerfault, C. Reina, D. Kafetzopoulos, G. Potamias, F. Ricoul, M. Tsiknakis, J. Schenkenburger, et P. Grangeat, « LOCCANDIA Lab-on-Chip Based Protein Profiling for Cancer Diagnosis, » dans *5th International Workshop on Wearable, Micro and Nanosystems for personalized Health (pHealth)*, 2008, page 21–23.

[6] P. Grangeat *et coll.*, « Final Project Review, technology integration, » LOCCANDIA, mars 2010, *CEA internal chrono number DTBS/STD/10-35*.

[7] R. Pérenon, A. Mohammad-Djafari, L. Duraffourg, et P. Grangeat, « Quantification moléculaire par spectrométrie de masse à base de NEMS : modélisation et inversion du problème. » dans *XXIIIème colloque GRETSI*, Bordeaux, France, 2011.

[8] R. Pérenon, A. Mohammad-Djafari, E. Sage, L. Duraffourg, S. Hentz, A. Brenac, R. Morel, et P. Grangeat, « MCMC-Based bayesian estimation algorithm dedicated to NEMS Mass Spectrometry, » dans *32nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science Engineering*, Garching, Germany, 2012.

[9] Consortium du projet, « Résumé du projet BHI-PRO, » 2010, contrat ANR-2010-BLAN-0313.

[10] R. Schiess, « Proteomic Strategy for Biomarker Discovery, » thèse de doctorat, ETH Zürich, Switzerland, 2008.

[11] M. Protter, M. Elad, H. Takeda, et P. Milanfar, « Generalizing the Nonlocal-Means to Super-Resolution Reconstruction, » *IEEE Transactions on Image Processing*, vol. 18, no. 1, pages 36–51, janvier 2009.

[12] J. Idier, éditeur, *Bayesian Approach to Inverse Problems*. London : ISTE Ltd and John Wiley & Sons Inc., 2008.

[13] P. Grangeat, éditeur, *Tomography*, 1ère édition. ISTE Ltd and John Wiley & Sons Inc., octobre 2009.

[14] K.-A. Do, P. Müller, et M. Vannucci, *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, 2006.

[15] J. Zhang, X. Zhou, H. Wang, A. Suffredini, L. Zhang, Y. Huang, et S. Wong, « Bayesian Peptide Peak Detection for High Resolution TOF Mass Spectrometry, » *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pages 5883–5894, novembre 2010.

[16] K. Harris, M. Girolami, et H. Mischak, « Definition of Valid Proteomic Biomarkers : A Bayesian Solution, » dans *PRIB '09 : Proceedings of the 4th IAPR International Conference on Pattern Recognition in Bioinformatics*. Berlin, Heidelberg : Springer-Verlag, 2009, page 137–149.

[17] M. Bhattacharjee, C. Botting, et M. Sillanpää, « Bayesian biomarker identification based on marker-expression proteomics data, » *Genomics*, vol. 92, no. 6, pages 384–392, décembre 2008.

[18] G. Strubel, J.-F. Giovannelli, C. Paulus, L. Gerfault, et P. Grangeat, « Bayesian estimation for molecular profile reconstruction in proteomics based on liquid chromatography and mass spectrometry, » dans *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Lyon, France, août 2007, pages 5979–5982.

[19] H. Jeffreys, *Theory of Probability*, 3ème édition. Clarendon Press, Oxford University Press, 1961.

[20] J. Albert, *Bayesian computation with R*. Springer Verlag, 2009.

[21] C. Robert et G. Casella, *Monte Carlo Statistical Methods*, 2ème édition. New York : Springer, 2004.

[22] M. J. Schervish, *Theory of Statistics*, 2ème édition. Springer, 1996.

[23] J. A. Hartigan, *Bayes Theory*. New York : Springer, 1983.

[24] C. Asseburg, « An Introduction to Using WinBUGS for Cost-Effectiveness Analyses in Health Economics (Part 1 : Bayesian Statistics : An Introduction), » Centre for Health Economics, University of York, UK, 2007.

[25] J. Kaipio et E. Somersalo, *Statistical and Computational Inverse Problems*, dans la série *Applied Mathematical Sciences*. New York : Springer, 2005, no. 160.

[26] A. C. Doyle, *A Study in Scarlet*, 1887.

[27] T. Bayes, « An essay towards solving a Problem in the Doctrine of Chances, » 1763. http://www.stat.ucla.edu/history/essay.pdf

[28] P.-S. Laplace, « Mémoire sur la probabilité des causes par les événements, » dans *Œuvres complètes de Laplace*. Académie des Sciences, 1774, vol. Tome VIII, pages 27–65.

[29] J. Skilling, « Nested sampling for general Bayesian computation, » *Bayesian Analysis*, vol. 1, no. 4, page 833–859, 2006.

[30] Z. Wang, R. M. Hope, Z. Wang, Q. Ji, et W. D. Gray, « Cross-subject workload classification with a hierarchical Bayes model, » *NeuroImage*, vol. 59, no. 1, pages 64–69, janvier 2012.

[31] A. Syversveen, « Noninformative Bayesian Priors. Interpretation And Problems With Construction And Applications. » *Preprint Statistics*, vol. 3, 1998.

[32] R. Kass et L. Wasserman, « The selection of prior distributions by formal rules, » *Journal of the American Statistical Association*, pages 1343–1370, 1996.

[33] P.-S. Laplace, *Essai philosophique sur les probabilités*. Bachelier, 1840.

[34] H. Jeffreys, « An Invariant Form for the Prior Probability in Estimation Problems, » *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pages 453–461, septembre 1946.

[35] J. Hartigan, « Invariant Prior Distributions, » *The Annals of Mathematical Statistics*, vol. 35, no. 2, pages 836–845, juin 1964.

[36] L. Held, *Methoden der statistischen Inferenz*. Springer Verlag, 2008.

[37] M. L. Eaton et W. D. Sudderth, « Invariance of posterior distributions under reparametrization, » *Sankhya A*, vol. 72, no. 1, pages 101–118, juin 2010.

[38] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.

[39] B. Hill, « Inference about variance components in the one-way model, » *Journal of the American Statistical Association*, page 806–825, 1965.

[40] E. L. Lehmann et G. Casella, *Theory of Point Estimation*. Springer, août 1998.

[41] Ghazali, « Le tabernacle des lumières, » Paris, 1994/env. 1100, الغزالي، مشكاة الأنوار.

[42] K. Strimmer, *Statistical Thinking : Introduction to Probabilistic Data Analysis*. en cours de rédaction, novembre 2010.

[43] G. Gan, C. Ma, et J. Wu, *Data Clustering : Theory, Algorithms, and Applications*. Society for Industrial Mathematics, juillet 2007.

[44] L. Devroye, « Non-uniform random variate generation, » dans *Simulation*, dans la série *Handbooks in Operations Research and Management Science*, 2006, no. 13, pages 83–121.

[45] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, et E. Teller, « Equation of state calculations by fast computing machines, » *The Journal of Chemical Physics*, vol. 21, no. 6, pages 1087–1092, 1953.

[46] W. Hastings, « Monte Carlo Sampling Methods Using Markov Chains And Their Applications, » *Biometrika*, vol. 57, no. 1, pages 97–109, 1970.

[47] C. Vacar, J.-F. Giovannelli, et Y. Berthoumieu, « Langevin and Hessian with Fisher approximation : Stochastic sampling for parameter estimation of structured covariance, » dans *ICASSP*, Prague, Czech Republic, mai 2011.

[48] J. Hobert et G. Casella, « The effect of improper priors on Gibbs sampling in hierarchical linear mixed models, » *Journal of the American Statistical Association*, vol. 91, no. 436, pages 1461–1473, 1996.

[49] R. Kass et A. Raftery, « Bayes factors, » *Journal of the american statistical association*, page 773–795, 1995.

[50] V. Smidl et A. Quinn, *The Variational Bayes Method in Signal Processing*. Springer-Verlag Berlin and Heidelberg GmbH & Co. K, novembre 2005.

[51] A. Fraysse et T. Rodet, « A gradient-like variational Bayesian algorithm, » dans *2011 IEEE Workshop on Statistical Signal Processing*. IEEE, juin 2011, pages 605–608.

[52] ——, « Sur un nouvel algorithme bayésien variationnel, » dans *XXIIIème Colloque GRETSI*. GRETSI, septembre 2011.

[53] M. J. Beal, « Variational algorithms for approximate Bayesian inference, » thèse de doctorat, Gatsby Computational Neuroscience Unit, University College London, 2003.

[54] A. Mohammad-Djafari, « A variational Bayesian algorithm for inverse problem of computed tomography, » dans *Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy (IMRT)*, Y. Censor, M. Jiang, et A. K. Louis, éditeurs. Edizioni Della Normale (CRM Series), 2008, pages 231–252.

[55] L. Chaari, F. Forbes, P. Ciuciu, T. Vincent, et M. Doiat, « Bayesian variational approximation for the joint detection estimation of brain activity in fMRI, » dans *IEEE Workshop on Statistical Signal Processing Proceedings*, 2011, pages 469–472.

[56] G. S. Omenn, D. J. States, M. Adamski, T. W. Blackwell, R. Menon, H. Hermjakob, R. Apweiler, B. B. Haab, R. J. Simpson, J. S. Eddes, E. A. Kapp, R. L. Moritz, D. W. Chan, A. J. Rai, A. Admon, R. Aebersold, J. Eng, W. S. Hancock, S. A. Hefta, H. Meyer, Y.-K. Paik, J.-S. Yoo, P. Ping, J. Pounds, J. Adkins, X. Qian, R. Wang, V. Wasinger, C. Y. Wu, X. Zhao, R. Zeng, A. Archakov, A. Tsugita, I. Beer, A. Pandey, M. Pisano, P. Andrews, H. Tammen, D. W. Speicher, et S. M. Hanash, « Overview of the HUPO Plasma Proteome Project : results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database, » *Proteomics*, vol. 5, no. 13, pages 3226–3245, août 2005.

[57] M. P. Murphy et H. LeVine, « Alzheimer's Disease and the β-Amyloid Peptide, » *Journal of Alzheimer's disease : JAD*, vol. 19, no. 1, page 311, janvier 2010.

[58] M. Trauchessec, « Compte rendu d'activité final du projet CAPSI, » Laboratoire d'Etude de la dynamique des protéomes, Grenoble, France, rapport technique, septembre 2009.

[59] V. Brun, C. Masselon, J. Garin, et A. Dupuis, « Isotope dilution strategies for absolute quantitative proteomics, » *Journal of Proteomics*, vol. 72, no. 5, pages 740–749, juillet 2009.

[60] B. Domon et R. Aebersold, « Mass Spectrometry and Protein Analysis, » *Science*, vol. 312, no. 5771, pages 212–217, avril 2006.

[61] E. J. Finehout, J. R. Cantor, et K. H. Lee, « Kinetic characterization of sequencing grade modified trypsin, » *PROTEOMICS*, vol. 5, no. 9, pages 2319–2321, juin 2005.

[62] T. Fortin, « Préparation d'échantillons, » Séminaire de spécification du projet BHI-PRO, janvier 2011.

[63] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, et A. Bairoch, « Protein Identification and Analysis Tools on the ExPASy Server, » dans *The Proteomics Protocols Handbook*, J. M. Walker, éditeur. Humana Press, 2005, pages 571–607.

[64] P. Artimo, M. Jonnalagedda, K. Arnold, D. Baratin, G. Csardi, E. de Castro, S. Duvaud, V. Flegel, A. Fortier, E. Gasteiger, A. Grosdidier, C. Hernandez, V. Ioannidis, D. Kuznetsov, R. Liechti, S. Moretti, K. Mostaguir, N. Redaschi, G. Rossier, I. Xenarios, et H. Stockinger, « ExPASy : SIB bioinformatics resource portal, » *Nucleic Acids Research*, vol. 40, no. W1, pages W597–W603, mai 2012.

[65] A. Felinger, *Data Analysis and Signal Processing in Chromatography*. Elsevier Science, mai 1998.

[66] G. Guiochon, A. Felinger, et D. G. G. Shirazi, *Fundamentals of Preparative and Nonlinear Chromatography*, 2ème édition. Academic Press, février 2006.

[67] Z. Pápai et T. L. Pap, « Analysis of peak asymmetry in chromatography, » *Journal of Chromatography A*, vol. 953, no. 1, page 31–38, 2002.

[68] V. Lange, P. Picotti, B. Domon, et R. Aebersold, « Selected reaction monitoring for quantitative proteomics : a tutorial, » *Molecular Systems Biology*, vol. 4, page 222, octobre 2008.

[69] S. Gallien, E. Duriez, et B. Domon, « Selected reaction monitoring applied to proteomics, » *Journal of Mass Spectrometry*, vol. 46, no. 3, page 298–312, 2011.

[70] P. Grangeat, C. Paulus, L. Gerfault, V. Kritsotakis, M. Tsiknakis, F. Lisacek, P. Binz, M. M. Perez-Perez, M. Trauchessec, et V. Brun, « First demonstration on NSE biomarker of a computational environment dedicated to lab-on-chip based cancer diagnosis, » dans *Proceedings of the 58th ASMS International Conference*, Salt Lake City, Utah, USA, 2010.

[71] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Rompp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.-A. Binz, et E. W. Deutsch, « mzML – a Community Standard for Mass Spectrometry Data, » *Molecular & Cellular Proteomics*, vol. 10, no. 1, pages R110.000 133–R110.000 133, août 2010.

[72] R. J. Carroll, D. Ruppert, L. A. Stefanski, et C. M. Crainiceanu, *Measurement Error in Nonlinear Models : A Modern Perspective, Second Edition*. Taylor & Francis, juin 2006.

[73] O. V. Krokhin et V. Spicer, « Predicting Peptide Retention Times for Proteomics, » dans *Current Protocols in Bioinformatics*, A. D. Baxevanis, G. A. Petsko, L. D. Stein, et G. D. Stormo, éditeurs. Hoboken, NJ, USA : John Wiley & Sons, Inc., septembre 2010.

[74] H. Choi, D. Fermin, et A. I. Nesvizhskii, « Significance Analysis of Spectral Count Data in Label-free Shotgun Proteomics, » *Molecular & Cellular Proteomics*, vol. 7, no. 12, pages 2373–2385, juillet 2008.

[75] W. M. Old, K. Meyer-Arendt, L. Aveline-Wolf, K. G. Pierce, A. Mendoza, J. R. Sevinsky, K. A. Resing, et N. G. Ahn, « Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics, » *Molecular & Cellular Proteomics*, vol. 4, no. 10, pages 1487 –1502, octobre 2005.

[76] C. Vogel et E. M. Marcotte, « Label-Free Protein Quantitation Using Weighted Spectral Counting, » dans *Quantitative Methods in Proteomics*, K. Marcus, éditeur. Totowa, NJ : Humana Press, 2012, vol. 893, pages 321–341.

[77] G. Wang, W. W. Wu, W. Zeng, C.-L. Chou, et R.-F. Shen, « Label-Free Protein Quantification Using LC-Coupled Ion Trap or FT Mass Spectrometry : Reproducibility, Linearity, and Application with Complex Proteomes, » *J. Proteome Res.*, vol. 5, no. 5, pages 1214–1223, octobre 2011.

[78] W. Zhu, J. Smith, et C. Huang, « Mass spectrometry-based label-free quantitative proteomics, » *J Biomed Biotechnol*, vol. 840518, 2010.

[79] L. Gerfault, G. Strubel, C. Paulus, J.-F. Giovannelli, et P. Grangeat, « Evaluation statistique d'un algorithme bayésien pour la reconstruction de profils moléculaires par spectrométrie de masse, » dans *XXIIème Colloque GRETSI*, Dijon, France, 2009.

[80] A. Barachant, S. Bonnet, M. Congedo, et C. Jutten, « Multiclass Brain Computer Interface Classification by Riemannian Geometry, » *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pages 920–928, avril 2012.

[81] M. Guindani, K. A. Do, P. Müller, et J. S. Morris, « Bayesian Mixture Models for Gene Expression and Protein Profiles, » dans *Bayesian inference for gene expression and proteomics*. Cambridge University Press, 2006, pages 238–253.

[82] N. Parrish, M. R. Gupta, et H. S. Anderson, « Robust classification of signal estimates given a channel model, » dans *2011 Workshop IEEE on Statistical Signal Processing*, Nice, France, 2011.

[83] I. Kuselman, F. Pennecchi, C. Burns, A. Fajgelj, et P. Zorzi, « Investigating out-of-specification test results of chemical composition based on metrological concepts, » *Accreditation and Quality Assurance*, vol. 15, pages 283–288, novembre 2009.

[84] T. Hastie, R. Tibshirani, et J. Friedman, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*, 2ème édition. Springer New York, 2009.

[85] F. Adjed, « Classification, apprentissage et sélection de modèles pour un mélange de populations appliqués en protéomique, » IMS, Bordeaux, Rapport de stage de Master 2, 2012.

[86] F. Lindsten, H. Ohlsson, et L. Ljung, « Clustering using sum-of-norms regularization : With application to particle filter output computation, » dans *2011 IEEE Workshop on Statistical Signal Processing*, 2011, page 201–204.

[87] P. P. Wang, D. Ruan, et E. E. Kerre, *Fuzzy Logic : A Spectrum of Theoretical & Practical Issues*. Springer-Verlag, 2007.

[88] M. Rantalainen et C. C. Holmes, « Accounting for control mislabelling in case-control biomarker studies, » *Journal of Proteome Research*, pages 5562–5567, octobre 2011.

[89] W. Förstner et B. Moonen, « A metric for covariance matrices, » *Qua vadis geodesia*, vol. 66, pages 113–128, 1999.

[90] J.-F. Giovannelli, P. Szacherski, et A. Giremus, « Découverte, Sélection, Apprentissage, Classification : autres racontars bayésiens. [Document de travail BHI-PRO], » avril 2012.

[91] J.-M. Marin et C. Robert, *Bayesian Core : A Practical Approach to Computational Bayesian Statistics*, dans la série *Springer Texts in Statistics*. New York : Springer, 2007.

[92] P. J. Green, « Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, » *Biometrika*, vol. 82, pages 711–732, 1995.

[93] T. Yu et H. Peng, « Quantification and deconvolution of asymmetric LC-MS peaks using the bi-Gaussian mixture model and statistical model selection, » *BMC Bioinformatics*, vol. 11, page 559, 2010.

[94] E. Grushka, « Characterization of exponentially modified Gaussian peaks in chromatography, » *Analytical Chemistry*, vol. 44, no. 11, pages 1733–1738, 1972.

[95] P. J. Naish et S. Hartwell, « Exponentially Modified Gaussian functions - A good model for chromatographic peaks in isocratic HPLC ? » *Chromatographia*, vol. 26, no. 1, pages 285–296, 1988.

[96] J. Listgarten, R. M. Neal, S. T. Roweis, P. Wong, et A. Emili, « Difference detection in LC-MS data for protein biomarker discovery, » *Bioinformatics*, vol. 23, no. 2, pages e198–e204, janvier 2007.

[97] R. O'Hara et M. Sillanpää, « A Review of Bayesian Variable Selection Methods : What, How and Which, » *Bayesian Analysis*, vol. 4, pages 85–118, 2009.

[98] A. Ndoye, « Sélection de variables appliquée à l'analyse protéomique par spectrométrie de masse en mode SRM, » CEA Leti, Université de Montpellier, Rapport de stage de Master 2, septembre 2012.

[99] J.-M. Marin, « On some computational methods for Bayesian model choice, » Rennes, 2008. http://mas2008.univ-rennes1.fr/download/papers/marin.pdf

Sites web

[100] {Proteome Software, Inc.}. Proteome software. 2012. http://www.proteomesoftware.com/Proteome_software_ed_proteomics.html.

[101] {mosaiques diagnostics and therapeutics AG}. Clinical proteomics. 2012. http://mosaiques-diagnostics.de/diapatpcms/mosaiquescms/front_content.php?idcat=166.

[102] J. Miller. Earliest known uses of some of the words of mathematics. 2012. http://jeff560.tripod.com/mathword.html.

[103] {Wikipedia contributors}. Conjugate prior, July 2012. http://en.wikipedia.org/w/index.php?title=Conjugate_prior&oldid=503475844. Page Version ID : 503475844.

[104] E. W. Weisstein. Correlation coefficient. 2012. http://mathworld.wolfram.com/CorrelationCoefficient.html.

[105] MathWorks. Statistical toolbox (v8.0) – {Matlab}. 2012.
http://www.mathworks.fr/products/statistics/.

[106] T. Minka. Lightspeed matlab toolbox. 2011.
http://research.microsoft.com/en-us/um/people/minka/software/lightspeed/.

[107] {Swiss Institute of Bioinformatics}. PeptideCutter. 2012.
http://web.expasy.org/peptide_cutter/.

[108] {Swiss Institute of Bioinformatics}. MSight, a new vision in mass spectrometry imaging (v3.0). 2012.
http://web.expasy.org/MSight/.

[109] MathWorks. Parallel computing toolbox (v6.0) – {Matlab}. 2012.
http://www.mathworks.fr/products/parallel-computing/.

[110] gp-you.org. GPUmat : GPU toolbox for Matlab (v0.280). June 2012.
http://gp-you.org/.

[111] M. Buehren. Multicore – parallel processing on multiple cores. September 2011.
http://www.mathworks.com/matlabcentral/fileexchange/13775.

[112] {MacCoss Lab Software}. Skyline (v1.3). 2012.
https://skyline.gs.washington.edu/labkey/project/home/software/Skyline/begin.view.

Bibliographie personnelle

Communications

[113] P. Szacherski, J.-F. Giovannelli, L. Gerfault, et P. Grangeat, « Apprentissage supervisé robuste de caractéristiques de classes. Application en protéomique, » dans *XXIIIème Colloque GRETSI*. Bordeaux, France : GRETSI 2011, septembre 2011.

[114] P. Szacherski, J.-F. Giovannelli, et P. Grangeat, « Reconstruction of proteomic profiles. Supervised learning of class characteristics, » Paris, France, novembre 2011, journée GdR ISIS : Methodes de Monte Carlo pour les problèmes inverses bayésiens.

[115] ——, « Joint Bayesian hierarchical inversion-classification and application in proteomics, » dans *2011 IEEE Workshop on Statistical Signal Processing*. IEEE, juin 2011, pages 121–124.

[116] P. Szacherski, J.-F. Giovannelli, L. Gerfault, A. Giremus, et P. Grangeat, « Robust MS serum sample classification in proteomics by the use of inverse problems, » dans *2012 IEEE International Workshop on Genomic Signal Processing and Statistics*, Washington, DC, USA, décembre 2012.

[117] L. Gerfault, P. Szacherski, J.-F. Giovannelli, J.-P. Charrier, P. Mahé, et P. Grangeat, « A hierarchical SRM acquisition chain model for improved protein quantification in serum samples, » dans *RECOMB-CP 2012*, San Diego, USA, avril 2012.

[118] G. Strubel, J.-F. Giovannelli, L. Gerfault, P. Szacherski, C. Paulus, M. Trauchessec, V. Brun, et P. Grangeat, « Bayesian Protein Quantification and Instrument Parameter Self-Calibration, » article non publié, août 2012.

[119] P. Grangeat, P. Szacherski, L. Gerfault, et J.-F. Giovannelli, « Bayesian hierarchical reconstruction of protein profiles including a digestion model, » dans *Proceedings of the 59th ASMS International Conference*, Denver, Colorado, USA, juin 2011.

[120] P. Szacherski, L. Gerfault, J.-F. Giovannelli, et P. Grangeat, « Quantification de protéines. Échantillonnage stochastique dans MATLAB, » Grenoble, France, novembre 2011.

[121] P. Szacherski, J.-F. Giovannelli, et P. Grangeat, « Inversion-Classification – Apport des problèmes inverses à la classification. Application en Protéomique, » Bordeaux, France, janvier 2012, séminaire d'équipe IMS/GSI.

[122] P. Szacherski, J.-F. Giovannelli, L. Gerfault, P. Mahé, J.-P. Charrier, A. Giremus, B. Lacroix, et P. Grangeat, « Classification of proteomic serum samples seen as hierarchical Bayesian inverse problem, » *to be submitted to IEEE/ACM Transactions on Computational Biology and Bioinformatics*, octobre 2012.

[123] L. Gerfault, P. Szacherski, J.-F. Giovannelli, J.-P. Charrier, P. Mahé, B. Lacroix, et P. Grangeat, « Présentation d'un algorithme d'inversion hiérarchique bayésien

pour la quantification et la classification de données protéomiques, » dans *Atelier Prospectom*, Grenoble, France, novembre 2012.

[124] F. Adjed, J.-F. Giovannelli, A. Giremus, P. Szacherski, C. Truntzer, N. Dridi, P. Roy, P. Ducoroy, V. Picaud, D. Maucort-Boulch, L. Gerfault, et P. Grangeat, « Vers la découverte et la sélection de marqueurs : une approche bayésienne, » dans *Atelier Prospectom*, Grenoble, France, novembre 2012.

Brevets

[125] P. Szacherski, J.-F. Giovannelli, et P. Grangeat, « Procédé et dispositif d'estimation de paramètres biologiques ou chimiques dans un échantillon, procédé d'aide au diagnostic correspondant, » FR Brevet 11 53 008, avril, 2011.

[126] ——, « Method and device for estimating biological or chemical parameters in a sample, corresponding method for aiding diagnosis, » US Brevet 13/438 977, avril, 2012.

Index

T

V

X

Reconstruction de profils protéiques pour la recherche de biomarqueurs

Résumé :

Cette thèse préparée au *CEA Léti*, Minatec Campus, Grenoble, et à l'*IMS*, Bordeaux, s'inscrit dans le thème du traitement de l'information pour des données protéomiques. Nous cherchons à reconstruire des profils protéiques à partir des données issues de chaînes d'analyse complexes associant chromatographie liquide et spectrométrie de masse. Or, les signaux cibles sont des mesures de traces peptidiques qui sont de faible niveau dans un environnement très complexe et perturbé. Ceci nous a conduits à étudier des outils statistiques adaptés. Ces perturbations peuvent provenir des instruments de mesure (variabilité technique) ou des individus (variabilité biologique). Le modèle hiérarchique de l'acquisition des données permet d'inclure ces variabilités explicitement dans la modélisation probabiliste directe. La mise en place d'une méthodologie *problèmes inverses* permet ensuite d'estimer les grandeurs d'intérêt. Dans cette thèse, nous avons étudié trois types de problèmes inverses associés aux opérations suivantes:

1. la quantification de protéines cibles, vue comme l'estimation de la concentration protéique,
2. l'apprentissage supervisé à partir d'une cohorte multi-classe, vu comme l'estimation des paramètres des classes, et
3. la classification à partir des connaissances sur les classes, vue comme l'estimation de la classe à laquelle appartient un nouvel échantillon.

La résolution des problèmes inverses se fait dans le cadre des méthodes statistiques bayésiennes, en ayant recours pour les calculs numériques aux méthodes d'échantillonnage stochastique (Monte Carlo Chaîne de Markov).

Mots-clés : problème inverse, modèles hiérarchiques, méthodes statistiques bayésiennes, MCMC, Gibbs, classification, apprentissage, quantification, protéomique, protéines, peptides, fragments, transitions, spectrométrie de masse, Full-MS, Selected Reaction Monitoring, chromatographie

Reconstruction of proteomic profiles for biomarker discovery

Abstract:

This thesis has been prepared at the *CEA Léti*, Minatec Campus, (Grenoble, France) and the *IMS* (Bordeaux, France) in the field of information and signal processing of proteomic data. The aim is to reconstruct the proteomic profile from the data provided by complex analytical workflow combining a spectrometer and a chromatograph. The signals are measurements of peptide traces which have low amplitude within a complex and noisy background. Therefore, adapted statistical signal processing methods are required. The uncertainty can be of technical nature (instruments, measurements) or of biological nature (individuals, "patients"). A hierarchical model, describing the forward problem of data acquisition, allows the explicit inclusion of those sources of variability within the probabilistic model. The use of the inverse problem methodology, finally, leads us to the estimation of the parameters of interest. In this thesis, we have studied three types of inverse problems for the following applications:

1. quantification of targeted proteins, seen as estimation of the protein concentration,
2. supervised training from a labelled cohort, seen as estimation of distribution parameters for each class,
3. classification given the knowledge about the classes, seen as estimation of the class a biological sample belongs to.

We solve these inverse problems within a Bayesian framework, resorting to stochastic sampling methods (Monte Carlo Markov Chain) for computation.

Keywords: inverse problem, hierarchical models, Bayesian statistical methods, MCMC, Gibbs, classification, statistical learning, quantification, proteomics, proteins, peptides, fragments, transitions, mass spectrometry, Full-MS, Selected Reaction Monitoring, chromatography

Commissariat à l'Énergie Atomique et aux Énergies Alternatives de Grenoble
Laboratoire d'Électronique et de Technologie de l'Information
MINATEC Campus
Département des micro-Technologies pour la Biologie et la Santé
Service Technologies pour la Détection
Laboratoire Électronique et Systèmes pour la Santé
17, rue des Martyrs
38054 GRENOBLE CÉDEX 9
FRANCE

Univ. Bordeaux
IMS
UMR 5218
351, cours de la Libération
F-33400 TALENCE
FRANCE